E-ISSN: 2717-4581 | P-ISSN: 2783-1345

Innovation Management and Operational Strategies

www.journal-imos.ir

Inov. Manage. Oper. Stra. Vol. 5, No. 3 (2024) 259-287.

Paper Type: Original Article

Deepfake Detection Models and Methods in Artificial Intelligence and Insights from Media and Social Culture Perspective

Sohail Fakheri[1,*], Azamossadat Nourbakhsh[1], Mohammadreza Yamaghani[1]

[1] *Department of Computer and Information Technology, Lahijan Branch, Islamic Azad University, Lahijan, Iran; fakherisoheil@iau.ac.ir; nourbakhsh@iau.ac.ir; o_yamaghani@iau.ac.ir.*

Citation:

Fakheri, S., Nourbakhsh, A., & Yamaghani, M. R. (2024). Deepfake detection models and methods in artificial intelligence and insights from media and social culture perspective. *Innovation management and operational strategies, 5*(3), 259-287.

Received: 17/04/2024 Reviewed: 20/06/2024 Revised: 10/07/2024 Accepted: 25/08/2024

Abstract

Purpose: This study explores the phenomenon of deepfakes as a consequence of rapid advancements in artificial intelligence, machine learning, and deep learning technologies over the past decade. The primary objective is to analyze various methods for detecting deepfakes and examine their social and legal implications.

Methodology: The research categorizes and evaluates four types of deepfake detection methods: deep learning-based, classical machine learning-based, statistical, and blockchain-based approaches. It also assesses the performance of these methods on different datasets.

Findings: The findings indicate that deep learning-based methods are more effective in detecting deepfakes compared to other approaches. Furthermore, the study analyzes the impact of deepfakes from multiple perspectives, including media and society, media production, representation, dissemination, audience, gender, law, and politics. The results reveal that society is currently unprepared to effectively combat deepfakes, due to a combination of technological, educational, and regulatory shortcomings.

Originality/Value: This research provides a comprehensive and comparative analysis of deepfake detection methods, offering valuable insights for policymakers and researchers. The study highlights the urgent need for effective strategies to address the rapidly evolving challenges posed by deepfakes in both social and legal contexts.

Keywords: Artificial Intelligence, Deepfake, Digital Media, Machine Learning, Deep Learning.

Corresponding Author: fakherisoheil@iau.ac.ir

10.22105/imos.2024.452298.1344

مدیریت نوآوری و راهبردهای عملیاتی
دوره ۵، شماره (۳)، (۱۴۰۳)، ۲۸۷-۲۵۹
www.journal-imos.ir

نوع مقاله: پژوهشی

مدل‌ها و روش‌های تشخیص دیپ‌فیک در هوش مصنوعی و تاثیر این پدیده بر فرهنگ اجتماعی از منظر رسانه

سهیل فاخری[۱،*]، اعظم‌السادات نوربخش[۱]، محمدرضا یمقانی[۱]
[۱]گروه کامپیوتر و فناوری اطلاعات، واحد لاهیجان، دانشگاه آزاد اسلامی، لاهیجان، ایران.

چکیده

هدف: این پژوهش به بررسی پدیده دیپ‌فیک به عنوان یکی از پیامدهای فناوری‌های پیشرفته هوش مصنوعی، یادگیری ماشین و یادگیری عمیق در دهه اخیر می‌پردازد. هدف این تحقیق، تحلیل روش‌های مختلف مقابله با محتوای جعلی و بررسی اثرات اجتماعی و قانونی آن‌ها است.

روش‌شناسی پژوهش: این مطالعه به بررسی چهار دسته از روش‌های مقابله با دیپ‌فیک شامل روش‌های مبتنی بر یادگیری عمیق، یادگیری ماشین کلاسیک، روش‌های آماری و بلاک‌چین می‌پردازد. همچنین، عملکرد این روش‌ها در شناسایی دیپ‌فیک‌ها بر اساس مجموعه‌داده‌های مختلف مورد ارزیابی قرار می‌گیرد.

یافته‌ها: نتایج نشان می‌دهند که روش‌های مبتنی بر یادگیری عمیق در شناسایی دیپ‌فیک‌ها کارایی بیشتری دارند. همچنین، این تحقیق به تحلیل جنبه‌های مختلف دیپ‌فیک از منظر رسانه‌ها، جامعه، تولید و بازنمایی رسانه‌ها، مخاطبان، جنسیت، قانون و سیاست می‌پردازد و نشان می‌دهد که جامعه در مقابله با دیپ‌فیک‌ها آماده نیست.

اصالت/ارزش افزوده علمی: این پژوهش با بررسی جامع و تطبیقی روش‌های مختلف شناسایی دیپ‌فیک، به ارزشمندی و اصالت خود در زمینه ارایه راهکارهای موثر و تحلیل ابعاد اجتماعی و قانونی این پدیده می‌افزاید. نتایج این مطالعه می‌تواند به سیاست‌گذاران و محققان در تدوین راهبردهای مناسب برای مقابله با دیپ‌فیک کمک کند.

کلیدواژه‌ها: هوش مصنوعی، دیپ فیک، رسانه دیجیتال، یادگیری ماشین، یادگیری عمیق.

۱- مقدمه

پیشرفت‌های قابل توجه در فناوری‌های مبتنی بر هوش مصنوعی[۱] و شبکه عصبی[۲] نقش اساسی در دستکاری محتوای چندرسانه‌ای ایفا می‌کند. به‌عنوان‌مثال، ابزارهای نرم‌افزاری مجهز به هوش مصنوعی مانند *Face App*[۳] و *Fake App*[۴] برای تعویض چهره در تصاویر و ویدیوها استفاده می‌شود. این امکان تعویض به هر کسی اجازه می‌دهد مدل مو، جنسیت، سن و سایر ویژگی‌های شخصی را در تصاویر و ویدئوها تغییر دهد. انتشار این ویدئوهای جعلی باعث نگرانی‌های زیادی در سطوح مختلف می‌شود، این ویژگی به نام دیپ‌فیک[۵] معروف شده است.

[1] Artificial Intelligence (AI)
[2] Artificial Neural Networks (ANN)
[3] FaceApp. Accessed: Jan. 4, 2021. Available: https://www.faceapp.com/
[4] FakeApp. Accessed: Jan. 4, 2021. Available: https://www. fakeapp.org/
[5] Deep fake

اصطلاح *Deepfake* از *Deep Learning* (*DL*) و *Fake* مشتق شده است و محتوای ویدیویی یا تصویری واقع‌گرایانه را توصیف می‌کند که با پشتیبانی یادگیری عمیق ایجاد شده‌اند. این کلمه از نام کاربری ناشناس در *Reddit* در اواخر سال ۲۰۱۷ آمده است که از روش‌های یادگیری عمیق برای جایگزینی تصویر یک شخص استفاده می‌کرد.

یکی از ویژگی‌های اساسی عصر ما تسلط رسانه‌های دیجیتال[1] بر جهات مختلف زندگی فردی، اجتماعی و شغلی انسان‌ها است، جایی که اطلاعات و محتوای دیجیتالی به راحتی می‌تواند ایجاد شود و در سطح جهانی منتشر گردد. درحالی‌که این امر دسترسی به اطلاعات را سهولت بخشیده، اما به این معنی است که تایید و اعتماد به چنین اطلاعاتی برای شهروندان به‌طور فزاینده‌ای چالش‌برانگیز شده است. پیشرفت‌های اخیر در هوش مصنوعی تاثیر عمیقی بر حوزه‌های مختلف، از جمله رسانه‌های دیجیتال داشته و پیامدهای قابل‌توجه‌ای را به دنبال خود آورده است. هوش مصنوعی به دلیل کاربردهای عملی چندگانه‌ای که در سال‌های اخیر داشته است، به‌عنوان یک فناوری تغییر پارادایم در نظر گرفته می‌شود که بیشتر آن‌ها به زیرشاخه‌ای از هوش مصنوعی به نام یادگیری عمیق[2] نسبت داده می‌شوند [1].

پیشرفت‌های اخیر در هوش مصنوعی، نمونه‌های متعددی از کاربردهای آن در زمینه‌های مختلف را به نمایش گذاشته است که نشان‌دهنده ارتباط گسترده و بین رشته‌ای آن است که شامل مواردی مثل توسعه خودروهای خودران، افزایش عملکرد انسانی، فراهم کردن تعاملات سطح انسانی در گفتار[3]، طراحی ربات‌های هوشمند، کشف و درمان بیماری‌ها، تولید تصاویر خلاقانه هنری، تولید و ویرایش متون، نقاشی‌ها و فیلم‌ها و مواردی از این دست می‌باشد. استفاده از قابلیت‌های پیشرفته یادگیری عمیق، هوش مصنوعی اکنون می‌تواند تصاویر، متون و صداها را با کیفیت و واقع‌گرایی بسیار بالا ایجاد یا تغییر دهد [2]. این قابلیت می‌تواند محتواهای تقلبی بسیار واقعی شامل متون، صداها، ویدئوها و عکس‌ها که در ابتدا ممکن است واقعی به نظر برسند، اما درواقع توسط الگوریتم‌های هوش مصنوعی تولید شده‌اند را تولید کند. در ادامه‌ی این مقاله به تفضیل به معرفی الگوریتم‌های هوش مصنوعی که در پژوهش‌های دیگر به آن‌ها پرداخته شده و الگوریتم‌های اصلی در تولید و تشخیص دیپ‌فیک به شمار می‌روند خواهیم پرداخت.

هوش مصنوعی می‌تواند با اضافه کردن تصویر صورت یک شخص به صورت کاراکتر اصلی در یک تصویر دیگر تصاویر تقلبی ایجاد کند و یا تغییر صدای شخصی را به‌طوری‌که اظهاراتی که هرگز واقعیت نداشته‌اند بیان می‌کند، نمایش دهد، همچنین ویدئوهای تقلبی بسیار واقع‌گرایانه ایجاد کند که تشخیص غیرواقعی بودن آن‌ها در نگاه اول بسیار دشوار خواهد بود یا توسط کاربر معمولی غیرممکن خواهد بود [4]–[2].

این پیشرفت‌های فناورانه نه تنها قادر به انقلاب در رسانه‌های دیجیتال هستند، بلکه پیامدهای مهمی از نظر اجتماعی نیز دارند. مشکلاتی که درباره نقض اعتماد عمومی به اصالت آنچه دیده، شنیده شده و بر باور عمومی تاثیر می‌گذارد که نیاز به توجه دقیق به تاثیرات اخلاقی و اجتماعی این تولیدات مبتنی بر هوش مصنوعی دارد. نیاز اصلی از منظر حقوق و قوانین پیرامون این مسایل نیز احساس می‌شود. همچنین بر اهمیت مطالعه دقیق درباره این تکنولوژی افزوده‌اند.

چندین فیلم دیپ‌فیک پربازدید در اینترنت وجود دارد که می‌توان آن‌ها را در وب‌سایت‌های معروفی مانند یوتیوب[4] پیدا کرد. یکی از اولین و احتمالا مشهورترین دیپ‌فیک‌ها، ویدیوی سال ۲۰۱۸ باراک اوباما است [5] که در آن او به طنز به خطرات دیپ‌فیک‌ها اشاره می‌کند و هشدار می‌دهد چیزی که درباره اوباما هرگز واقعیت نداشته و این‌چنین اظهاراتی نداشته است. علاوه بر این، دیپ‌فیک‌های مختلف دیگری نیز وجود دارند که برخی از آن‌ها مربوط به افراد سیاست‌مدار می‌باشد که با تعویض چهره آن‌ها را با چهره‌های کاملا متفاوتی نشان می‌دهند یا اظهارات این افراد را تغییر می‌دهند، به‌گونه‌ای که اظهارات یک شخص را با تصویر و صدای شخص دیگری به نمایش می‌گذارند. در این مقوله مساله‌ی انتشار اخبار جعلی در سطح جامعه نیز به وجود می‌آید که خود زمینه بسیاری از هرج‌ومرج‌ها یا گمراهی‌ها را ایجاد می‌کند. به‌عنوان‌مثال در ویدئویی چهره دونالد ترامپ با چهره مستر بین جایگزین شده است. همچنین برخی تصاویر یا ویدئوها که شامل فیلم‌های غیراخلاقی پورنوگرافی از بازیگران معروف است نیز موجود

[1] Digital media
[2] Deep Learning (DL)
[3] Human-level spoken interaction
[4] Youtube.com

است. مجموعه‌ی چنین مواردی باعث شده شناخت از نحوه ساخت، تاثیر آن بر افراد و جامعه و همچنین تشخیص محتوای دیپ‌فیک از واقعی بسیار حایز اهمیت باشد که درواقع دلیل اصلی نگارش این پژوهش است.

برای تولید چنین ویدئوهای تقلبی، از دو شبکه عصبی شامل یک شبکه مولد[1] و یک شبکه متمایز[2] با تکنیک *Face Swap* استفاده شده [5] که شبکه مولد با استفاده از رمزگذار و رمزگشا تصاویر جعلی را ایجاد می‌کند. شبکه متمایزکننده اصالت تصاویر تازه تولید شده را تعریف می‌کند. ترکیب این دو شبکه، شبکه‌های متخاصم مولد[3] (GAN) نامیده می‌شود که توسط ایان گودفلو پیشنهاد شده است [6].

ویدیوهای دیپ‌فیک به‌طور مداوم در شبکه‌های اجتماعی مختلف ایجاد و به اشتراک گذاشته می‌شوند و طیف گسترده‌ای از موضوعات را پوشش می‌دهند. در زمان نگارش این متن، دیپ‌فیک‌های معروفی از بازیگران معروف مانند مورگان فریمن، تام کروز، کیانو ریوز و... در شبکه‌های اجتماعی مختلف وجود دارد، همچنین نسخه‌های جایگزینی از برخی از قسمت‌های فیلم سینمایی هیجان‌انگیز و روانشناختی آمریکایی مشهور راننده تاکسی محصول سال ۱۹۷۶ نیز وجود دارد که در این نسخه تغییر یافته، چهره بازیگر اصلی، یعنی رابرت دنیرو، با چهره آل پاچینو به‌طور بی‌درنگ در تمام فیلم جایگزین شده است [7].

علاوه بر این، مواردی از ساخت آواتارهای بلادرنگ[4] برای سیستم‌های تله‌کنفرانسی وجود دارد تا فقط با استفاده از یک عکس به سادگی نمایه‌های دیپ‌فیک را به‌صورت بلادرنگ برای پلتفرم‌هایی مانند اسکایپ و زوم بسازند [8]. این مثال‌ها نشان می‌دهد که سرعت همه‌گیری فناوری دیپ‌فیک چقدر بالا است و چالش‌های قابل توجهی را در تشخیص حقیقت از دروغ در دوران دیجیتال مطرح می‌کنند. دیپ‌فیک‌ها نه تنها قابلیت اضافه کردن تصاویر به ویدیوها را دارند بلکه می‌توانند محتوای کاملا جدیدی نیز که شامل تصاویر بسیار واقع‌گرایانه از چهره‌های انسانی است [9]، تولید کنند که به‌وسیله یادگیری ماشینی از تصاویر انسان‌های دیگر به‌دست می‌آید [10]، یا اختلالی که توسط کاربران جعلی در مواقع حیاتی هنگام تله‌کنفرانس‌های تجاری برای عقد قراردادهای مالی ایجاد می‌شود را درنظر بگیرید، به‌ویژه در میان همه‌گیری کووید-۱۹ که استفاده از کنفرانس‌های مجازی رونق پیدا کرده بود که البته هنوز هم در جریان است.

بر اساس پژوهش پاترینی و همکاران [11]، موتور جستجوی گوگل می‌تواند صفحه وبی که حاوی ویدئوهای مرتبط با دیپ فیک هستند را پیدا کند که خلاصه‌ای از این گزارش در **شکل ۱** آمده است. این گزارش مربوط به سال ۲۰۱۸ می‌باشد و نکته حایز اهمیت رشد نمایی این آمار است. فقط در سال ۲۰۲۱ حدود ۲۰ هزار ویدئو و صوت با تکنولوژی دیپ‌فیک تولید شده است که برای سال ۲۰۲۴ با رشد نمایی پیش‌بینی عدد ۵۰۰ هزار می‌شود. تعداد دقیق سایت‌هایی که در یک سال خاص ویدئوهای دیپ‌فیک منتشر کرده‌اند، مشخص نیست. دلیل این امر آن است که بسیاری از این سایت‌ها به‌طور غیرقانونی و مخفیانه فعالیت می‌کنند و دایما در حال تغییر آدرس و نام خود هستند. بااین‌حال، برخی از منابع تخمین‌هایی را ارایه می‌دهند. به‌عنوان‌مثال، *Deep trace* یک شرکت امنیت سایبری، در گزارشی تخمین زده است که در سال ۲۰۲۲ بیش از یکصدهزار ویدئو دیپ‌فیک در ۱۰۰۰ سایت مختلف منتشر شده است. همچنین یک شرکت دیگر فعال در زمینه امنیت سایبری، در گزارشی دیگر تخمین زده است که در سال ۲۰۲۲ بیش از پانصدهزار ویدئو دیپ‌فیک در ۳۰۰۰ سایت مختلف منتشر شده است، اما نکاتی که باید در نظر داشت، از قرار زیر است:

۱. این تخمین‌ها فقط بر اساس تعداد سایت‌هایی است که شناسایی شده‌اند و ممکن است تعداد واقعی سایت‌های فعال در این زمینه بیشتر باشد.
۲. بسیاری از ویدئوهای دیپ‌فیک در شبکه‌های اجتماعی و پلتفرم‌های اشتراک‌گذاری ویدئو مانند یوتیوب و اینستاگرام منتشر می‌شوند.
۳. ردیابی و شمارش دقیق تعداد ویدئوهای دیپ‌فیک منتشر شده در این پلتفرم‌ها دشوار است.

[1] Generative network
[2] Discriminative network
[3] Generative Adversarial Networks (GAN)
[4] Real time

شکل ۱- سمت چپ آبی: نتایج موتور جستجوی گوگل تعداد صفحات وب حاوی کلیدواژه "Deepfake" (تعداد صفحات وب در برابر ماه). سمت راست سبز: موتور جستجوی گوگل صفحات وب حاوی ویدیوهای مرتبط با کلیدواژه "Deepfake" (شمار صفحات وب در برابر ماه).

Figure 1 - Blue left side: Google search engine results number of web pages containing the keyword "Deepfake" (number of web pages vs. month). Green right: Google search engine web pages containing videos related to the keyword "Deepfake" (number of web pages per month).

متاسفانه فرآیند کنونی برای مقابله با عدم وقوع دیپ‌فیک‌ها همچنان بهترین حالت را ندارند. با وجود پیچیدگی فنی فناوری مبتنی بر آن، ملزومات ایجاد دیپ‌فیک بسیار ساده است، با ابزارهای دسترسی‌پذیر برای عموم قادر به ساخت دیپ‌فیک خواهیم بود [11] که معمولا شامل کامپیوترهای استاندارد خانگی با کارت‌های گرافیک معمولی است. ساخت عکس و در برخی موارد ویدئو چیز جدیدی نیست، اما سهولت دستیابی به این امر، همراه با نتایج واقعی، چیزی است که جدید و هیجان‌انگیز اما از منظرهای مختلف نگران کننده است.

با توجه به تهدیدات حریم خصوصی، مطالعه‌ی دیپ‌فیک بسیار سریع آغاز شد. راسلر و همکاران [12] در مارس ۲۰۱۸ یک مجموعه داده ویدیو بی برای آموزش ابزارهای تشخیصی و تشخیص *Deepfake* به نام *FaceForensic* معرفی کردند. پس از یک ماه، محققان دانشگاه استنفورد روشی به نام "پرتره‌های ویدیو بی عمیق" منتشر کردند که عکس‌های واقعی را همانند ویدئو متحرک می‌کرد [13]. محققان دانشگاه کالیفرنیا برکلی رویکرد دیگری برای انتقال حرکات بدن یک فرد به فرد دیگری در این ویدئو ایجاد کردند و شرکت انو یدیا یک معماری مبتنی بر سبک[1] را برای *GANs* برای تولید تصو یر مصنوعی معرفی کرد [14].

مانند هر فن‌آوری جدید که نظرات متفاوتی پیرامون آن وجود دارد، اغلب افراد در این زمینه شروع به پیش‌بینی پیامدهای آینده آن بر افراد و جامعه کردند. بااین‌حال، دیدگاه‌های متفاوتی در این‌باره در رسانه‌ها مطرح می‌شود که تمایل دارند جنبه‌های متفاوتی از فناوری و پیامدهای آن را برای درک عموم آشکار سازند. در نتیجه، شایسته است به بررسی اینکه چگونه محتوای دیپ‌فیک را تشخیص دهیم و همچنین در حالت کلی هوش مصنوعی، کاربردها و اثرات آن در رسانه‌های معاصر چگونه به تصو یر کشیده می‌شود و مردم چگونه آن‌ها را چه به‌عنوان یک تهدید یا یک فرصت تلقی می‌کنند بپردازیم. رسانه‌های مختلف اغلب به جذابیت‌های موضوعات جدید مانند دیپ‌فیک می‌پردازند؛ به‌ویژه به دلیل توانایی آن در ایجاد محتوای متقاعدکننده و درعین‌حال ساختگی، توجه عمومی را به خود جلب کرده است و نگرانی‌هایی را در مورد اطلاعات نادرست و نقض حریم خصوصی ایجاد می‌کند. بااین‌حال، تشخیص این نکته ضروری است که فناوری دیپ فیک نیز کاربردهای قانونی دارد، مانند سرگرمی و هنر دیجیتال که پتانسیل آن را در زمینه طراحی و ایده‌پردازی برجسته می‌کند. به‌عنوان‌مثال زمانی که در آثار سینمایی قرار است چهره افراد مشهور پخش شود یا افرادی که دیگر در قید حیات نیستند، همچنین در جلوه‌های ویژه بصری برای پیر کردن بازیگر جوان، درحال حاضرکاربردهای این‌چنینی به‌صورت رایج در شرکت‌های بزرگ فیلم‌سازی انجام می‌پذیرد.

علاوه بر این درک عمومی از هوش مصنوعی و فناوری‌های مرتبط با آن، نه تنها تحت تاثیر تصو یر رسانه‌ای، بلکه تحت تاثیر تجربیات شخصی و هنجارهای اجتماعی نیز قرار می‌گیرد. درحالی‌که برخی ممکن است هوش مصنوعی را ابزاری برای نوآوری و پیشرفت بدانند، برخی دیگر ممکن

[1] Style-based generator architecture

است نگران تاثیر آن بر اشتغال، حریم خصوصی و امنیت باشند؛ بنابراین، پژوهش‌های اجتماعی درمورد هوش مصنوعی باید تلاش کند تا اطلاعات متعادل و دقیق را فراهم کند و بحث‌ها و تصمیم‌گیری‌های آگاهانه را بین مردم و دولت‌ها یا تصمیم‌گیرندگان تقویت کند. پیامدهای اخلاقی فناوری‌های هوش مصنوعی، از جمله دیپ‌فیک‌ها، مستلزم بررسی دقیق است. مسایلی مانند رضایت، اصالت و مسئولیت‌پذیری باید موردتوجه قرار گیرد تا از توسعه و استقرار مسئولانه این فناوری‌ها اطمینان حاصل شود. علاوه بر این، تلاش‌ها برای کاهش پیامدهای منفی هوش مصنوعی باید با ابتکاراتی برای به حداکثر رساندن سهم مثبت آن، مانند افزایش دسترسی و فراگیری در راه‌حل‌های مبتنی بر هوش مصنوعی تکمیل شود.

درنهایت، گفتمان پیرامون هوش مصنوعی و فناوری دیپ‌فیک منعکس‌کننده بحث‌های اجتماعی گسترده‌تر درباره نقش فناوری در شکل‌دهی به زندگی ما و نیاز به رویکردهای اخلاقی، شفاف و فراگیر برای نوآوری است. با درگیر شدن در گفتگوی سازنده و حل مشکلات مشترک، می‌توانیم از پتانسیل هوش مصنوعی برای منافع جمعی استفاده کنیم و درعین‌حال خطرات آن را کاهش دهیم.

رسانه‌های دیجیتال چندین ویژگی کلیدی دارند که بر ماهیت و نحوه استفاده از آن‌ها تاثیر می‌گذارد. اگرچه آن‌ها ادامه‌دهنده راه رسانه‌های سنتی هستند، اما قابلیت‌های آن‌ها انتشار اطلاعات را بسیار آسان می‌کند و به‌این‌ترتیب، مخاطبان هدف می‌توانند با حجم عظیم و بدون محدودیت مکانی یا زمانی محتوا مواجه شوند. استفاده از رسانه‌های دیجیتال به‌عنوان یک رسانه ارتباطی از طریق اینترنت، مخاطبان جهانی و تعامل را بیش‌ازپیش امکان‌پذیر کرده است، که قبلا در انقلاب‌های اجتماعی مانند بهار عربی یا حتی مسایل سیاسی انتخاباتی مورد استفاده قرار گرفته‌اند [15].

دیپ‌فیک به شدت با رسانه‌های دیجیتال و به‌ویژه رسانه‌های اجتماعی مرتبط است که از طریق آن به مخاطبان گسترده‌ای می‌رسد. ازآنجایی‌که متن، تصاویر، ویدئوها و صدا عناصر کلیدی تعامل و ارتباط در حوزه عمومی هستند، پیامدهای آن نیز تاثیر زیادی بر عموم جامعه دارد، درحالی‌که قدرت و قابلیت‌های هوش مصنوعی هنوز در حال رشد است، باید این نکته را در نظر بگیریم که به دور از قانون‌گذاری درست تحلیل‌های مفسرین در رسانه‌ها نیز با جهت‌گیری‌هایی ممکن است همراه باشد. تلاقی رسانه‌های دیجیتال و هوش مصنوعی بسیار موردتوجه است، زیرا سنگ‌بنای ارتباطات رسانه‌ای در عصر مدرن را با آخرین پیشرفت‌های فناوری در هوش مصنوعی پیوند می‌دهد که مرزهای واقعیت و انتشار اطلاعات را از بین می‌برد. با دیپ‌فیک‌ها، توانایی تحریف واقعیت یک جهش تصاعدی به جلو داشته است [16]؛ بنابراین، بررسی این تکنولوژی از دیدگاه‌های مختلف اهمیت زیادی دارد.

بااین‌حال، مروری جامع در این حوزه تحقیقاتی با ارایه اطلاعات خلاصه‌شده درباره دیپ‌فیک در همه جنبه‌ها، از جمله مجموعه داده‌ها و پژوهش‌های موجود برای محققان و متخصصان مفید خواهد بود. برای رسیدن به این هدف، ما یک مرور ادبیات سیستماتیک[1] در مورد تشخیص دیپ‌فیک در این مقاله ارایه می‌کنیم. هدف ما توصیف و تحلیل زمینه‌های مشترک و تنوع رویکردها در شیوه‌های فعلی در تشخیص دیپ‌فیک است. مطالب آتی در این پژوهش ما به شرح زیر خلاصه می‌شود:

ما ابزارها، تکنیک‌ها و مجموعه داده‌های موجود را برای تحقیقات مربوط به تشخیص دیپ‌فیک با طرح برخی سوالات تحقیقاتی گزارش می‌کنیم، همچنین یک طبقه‌بندی معرفی می‌کنیم که تکنیک‌های تشخیص دیپ‌فیک را در چهار دسته کلی با مروری بر دسته‌های و ویژگی‌های آن‌ها طبقه‌بندی می‌کند که این یک دسته‌بندی منظم و جدید است. ما یک تجزیه‌وتحلیل عمیق از شواهد تجربی مطالعات اولیه انجام می‌دهیم. همچنین، ما عملکرد روش‌های مختلف تشخیص دیپ‌فیک را با استفاده از معیارهای اندازه‌گیری مختلف ارزیابی می‌کنیم که در ادامه چند منبع مطالعاتی مفید را برجسته و معرفی می‌کنیم و دستورالعمل‌هایی را در مورد تشخیص دیپ‌فیک ارایه می‌دهیم که ممکن است به تحقیقات آتی در این زمینه کمک کند.

ادامه شرح مقاله به ترتیب زیر سازمان‌دهی شده است: بخش دوم روش بررسی را با تعریف سوالات تحقیق ارایه می‌دهد. در بخش سوم، ما به‌طور کامل یافته‌های مطالعات مختلف را موردبحث قرار می‌دهیم. بخش چهارم مشاهدات کلی مطالعه را بیان می‌کند و ما چالش‌ها و محدودیت‌ها را در بخش پنجم ارایه می‌کنیم. درنهایت، در بخش نهایی پیشنهادات آتی، بحث و نتیجه‌گیری ارایه می‌گردد.

[1] Systematic Literature Review (SLR)

۲- شرح مطالعات

در این مطالعه درمجموع ۱۱۲ پژوهش را از منابع معتبر در طی سه سال از دوره انتشار جمع‌آوری کردیم. دوره انتشار تحقیقات مربوط به دیپ‌فیک عمدتا در سال ۲۰۱۸ ظاهر شد؛ بنابراین، دوره انتشار را از ابتدای سال ۲۰۱۸ درنظر گرفتیم. در طول این بازه، تعداد انتشارات به‌طور تصاعدی افزایش یافته است. بر اساس تخمین‌های مختلف، می‌توان گفت که در سال ۲۰۲۱ حدود ۳۰۰۰ تا ۴۰۰۰ مقاله در مورد دیپ‌فیک منتشر شده است. این عدد با جست‌وجو در پایگاه گوگل اسکالر در این سال ۳۵۰۰ بوده است. برای سال ۲۰۲۲ این عدد به حدود ۵۰۰۰ و برای سال ۲۰۲۳ به حدود ۸۵۰۰ می‌رسد. این اعداد با تقریبا ۱٫۵ برابر افزایش انتشارات سال جدید، که نشان‌دهنده عطش تحقیق در حوزه دیپ‌فیک است. در این پژوهش ما عمدتا هشت منبع انتشار مختلف را از کنفرانس‌ها، کارگاه‌ها، مجلات و آرشیوهای معتبر در نظر می‌گیریم.

۲-۱- سوال اول: تکنیک‌های رایج تشخیص دیپ‌فیک چیست؟

بررسی کلی را در قالب برخی از سوالات تحقیق بررسی می‌کنیم. به‌عنوان بخشی از بحث، ابتدا تکنیک‌های تشخیص دیپ‌فیک را که به‌طور گسترده در ادبیات استفاده می‌شود، تعیین می‌کنیم. اگرچه دیپ‌فیک عمدتا تصاویر یا ویدیوها را با استفاده از تکنیک مبتنی بر یادگیری عمیق (*DL*) دستکاری می‌کند، روش‌های دیگر همراه با *DL Deepfake* نیز وجود دارد. پژوهش‌های مختلف را با توجه به تکنیک‌های کاربردی دسته‌بندی می‌کنیم و در بخش‌های بعدی به تشریح آن‌ها می‌پردازیم.

۲-۱-۱- روش‌های مبتنی بر یادگیری ماشین

الگوریتم‌های یادگیری ماشین سنتی (*ML*) در درک منطق هر تصمیمی که می‌تواند با اصطلاحات انسانی بیان شود، مفید است. چنین روش‌هایی برای حوزه دیپ‌فیک مناسب هستند، زیرا درک بهتری از داده‌ها و فرآیندهای آن وجود دارد. علاوه بر این، تنظیم پارامترها و تغییر طرح‌های مدل قابل مدیریت است. رویکردهای یادگیری ماشین مبتنی بر درخت، به‌عنوان‌مثال، درخت تصمیم[1]، جنگل تصادفی[2] و ...، الگوریتم‌های مناسبی برای دسته‌بندی (*Classification*) و رگرسیون (*Regression*) هستند و فرآیند تصمیم‌گیری را در قالب یک فلوچارت گام‌به‌گام درختی نشان می‌دهند.

شبکه‌های مولد متخاصم یا به اختصار *GAN* برای آموزش خودکار یک مدل تولیدی با فرض مساله بدون نظارت و ایجاد چهره‌های جعلی عکس واقعی در تصاویر یا ویدیوها استفاده می‌شوند. برخی از روش‌های مبتنی بر یادگیری ماشین می‌خواهند برخی از بی‌نظمی‌های موجود در چنین *GAN* هایی را نشان دهند. یک رویکرد اساسی دیپ‌فیک دستکاری چهره انسان برای گیج کردن مخاطبان آن است. رویکردهای مختلفی برای انجام آن وجود دارد. بااین‌حال، بیشتر تکنیک‌ها برای فریب دادن کاربران، نواحی خاصی از صورت را ویرایش می‌کنند، مانند سایه چشم، گوش با حلقه و ریش و سیبیل، رنگ مو و .... چنین ویرایش‌هایی با استفاده از یک قسمت محدود به شناسایی یا تشخیص ناحیه دستکاری شده است. برای غلبه و یا شناسایی بر این موارد، نویسندگان در پژوهش ماترن و همکاران [17] یک تکنیک دیپ‌فیک را با ترکیب مجموعه‌ای از این ویژگی‌ها پیشنهاد کردند.

در پژوهش سیفتسی و همکاران [18]، شدت علایم بیولوژیکی همراه با جهت‌های مکانی و زمانی [21]–[19] اندازه‌گیری می‌شود تا از نقاط شاخص مختلف صورت به‌عنوان‌مثال چشم، بینی و دهان استفاده شود؛ به‌عنوان ویژگی‌های منحصربه‌فرد برای احراز هویت ویدیوها یا تصاویر تولید شده توسط *GAN*. ویژگی‌های مشابهی در ویدیوهای دیپ‌فیک نیز قابل مشاهده است که می‌توان آن‌ها را با تقریب حالت سه‌بعدی سر کشف کرد [22]. در بیشتر موارد، حالات صورت در ابتدا با حرکات سر مرتبط است. حبیبا و همکاران [23] از شبکه پرسپترون چندلایه[3] *MLP* برای تشخیص ویدیوی دیپ‌فیک با قدرت محاسباتی کم با بهره‌گیری از ویژگی‌های مصنوعی بصری در ناحیه صورت استفاده کردند. با توجه به نگرانی عملکرد در روش‌های دیپ‌فیک مبتنی بر یادگیری ماشین، مشاهده می‌شود که این رویکردها می‌توانند تا ۹۸% دقت در تشخیص دیپ‌فیک به دست آورند. بااین‌حال، کیفیت عملکرد الگوریتم‌ها برای تشخیص کاملا به نوع مجموعه داده‌ها، ویژگی‌های انتخاب‌شده و هم‌ترازی بین داده‌های تمرین و مجموعه‌های آزمایشی بستگی دارد. زمانی که آزمایش از یک مجموعه داده مشابه با تقسیم آن به سطح معینی از نسبت، به‌عنوان‌مثال، ۸۰% برای یک مجموعه

[1] Decision tree
[2] Random forest
[3] Multilayer perceptron

تمرین و ۲۰% برای یک مجموعه آزمایشی، از یک مجموعه داده مشابه استفاده می‌کند، مطالعه می‌تواند نتیجه بالاتری به دست آورد. مجموعه داده‌های نامرتبط عملکرد را نزدیک به ۵۰% کاهش می‌دهد.

۲-۱-۲- روش‌های مبتنی بر یادگیری عمیق

در مورد تشخیص دیپ‌فیک در تصاویر، پژوهش‌های زیادی وجود دارد که روش‌های مبتنی بر یادگیری عمیق را برای شناسایی تصاویر جعلی تولید شده استفاده می‌کنند. ژانگ و همکاران [24] یک شبیه‌ساز *GAN* را معرفی کردند که جعل‌های تصویر *GAN* جمعی را تکرار می‌کند و آن‌ها را به‌عنوان ورودی به طبقه‌بندی کننده می‌دهد تا آن‌ها را به‌عنوان دیپ‌فیک شناسایی کند. ژو و همکاران [25] شبکه‌ای را برای استخراج ویژگی‌های استاندارد از داده‌های *RGB* پیشنهاد کردند، درحالی‌که چو وهمکاران [26] کاری مشابه اما عمومی را پیشنهاد کردند. علاوه بر این، در پژوهش‌های [29]–[27]، محققان چارچوب تشخیص جدیدی را بر اساس اندازه‌گیری فیزیولوژیکی، به‌عنوان‌مثال، ضربان قلب و تعداد پلک‌ها را پیشنهاد کردند. در ابتدا، روش مبتنی بر یادگیری عمیق در [30] برای تشخیص ویدیوی دیپ‌فیک پیشنهاد شد. دو ماژول اولیه، *Meso-4* و *MesoInception-4*، برای ساخت شبکه پیشنهادی خود استفاده شد. در این تکنیک، خطای میانگین مربعات[1] بین برچسب‌های واقعی و مورد انتظار به‌عنوان تابع ضرر[2] برای آموزش استفاده می‌شود. بهبود *Meso-4* در [31] پیشنهاد شده است.

در یک سناریوی نظارت شده [32]، نویسندگان نشان می‌دهند که *CNNs* عمیق [35]–[33] از *CNNs* کم‌عمق بهتر عمل می‌کنند. برخی از روش‌ها برای استخراج ویژگی‌های جعلی [36]، [37]، ویژگی‌های فضایی-زمانی [41]–[38]، بافت‌های مشترک صورت درمجموعه داده‌ها [42]، [43] و ۶۸ نقطه شاخص چهره [46]–[44] همراه با مصنوعات بصری (مانند چشم، دندان، حرکت لب و ...) از فریم‌های ویدیو استفاده می‌کنند.

چنین ویژگی‌هایی به‌عنوان ورودی این شبکه‌ها برای تشخیص دستکاری‌ها (دیپ‌فیک) استفاده شد. علاوه بر افزایش داده‌ها [47]، بازسازی وضوح فوق‌العاده [48] استراتژی‌های محلی‌سازی در سطوح پیکسل [12] بر روی کل فریم فرموله می‌شوند و حداکثر میانگین اختلاف (*MMD*) [49] برای کشف یک ویژگی کلی‌تر اعمال می‌شود. نوآوری‌های بیشتر با معرفی مکانیسم توجه به دست می‌آید [50]. درحالی‌که نتایج امیدوارکننده در پژوهش‌های [51]، [52] با استفاده از معماری به نام شبکه کپسول (*CN*) نشان داده شده است. *CN* نسبت به شبکه‌های بسیار عمیق به تعداد پارامترهای کمتری برای آموزش نیاز دارد. یک تکنیک یادگیری گروهی در پژوهش [53]، [54] برای افزایش عملکرد چنین ساختارهایی استفاده می‌شود که بیش از ۹۹% دقت را به دست می‌آورد.

مشاهده می‌کنیم که روش‌های زیادی برای اعمال تحلیل فریم به فریم در فیلم‌ها یا تصاویر برای دستکاری‌های صورت و ردیابی حرکت صورت برای به دست آوردن عملکرد بهتر پیشنهاد شده‌اند. برای مثال، در [60]–[55] شبکه‌های عصبی بازگشتی[3] برای استخراج ویژگی‌ها در سطوح مختلف میکرو و ماکروسکوپی برای تشخیص دیپ‌فیک پیشنهاد شده‌اند. صرف‌نظر از این نتایج هیجان‌انگیز در تشخیص، مشاهده می‌شود که بیشتر روش‌ها به بیش‌برازش متمایل هستند. تکنیک مبتنی بر جریان نوری[4] [61] و معماری‌های مبتنی بر رمزگذار خودکار [65]–[62] برای حل چنین مشکلاتی معرفی شده‌اند. یک ماسک مبتنی بر پیکسل [63] بر روی مدل‌های مختلف اعمال می‌شود تا تصویری ضروری از ناحیه آسیب‌دیده صورت به دست آید. فرناندو و همکاران [66] رویکردهای آموزشی خصمانه را به دنبال مکانیسم‌های مبتنی بر توجه برای دستکاری های پنهان صورت اعمال کرد. در [67]، محققان یک تکنیک خوشه‌بندی را با ادغام یک اصطلاح منظم سازی تعبیه سه‌گانه مبتنی بر حاشیه در طبقه‌بندی خود پیشنهاد کردند. درنهایت مساله طبقه‌بندی سه کلاسه را به مساله طبقه‌بندی دوکلاسه تبدیل کردند. لینچ و همکاران [68] یک تکنیک پیش‌پردازش داده را برای تشخیص دیپ‌فیک با استفاده از روش‌های *CNN* پیشنهاد کردند.

چیو و همکاران [69] شبکه‌های عصبی کانولوشنال پچ و جفت[5] را پیشنهاد کردند. دو و همکاران [70] با استفاده از ویژگی‌های الگوهای پنهان تصویر، تحلیلی را در حوزه فرکانس انجام دادند. یک رویکرد مدرن به نام *ID-revelation* برای یادگیری ویژگی‌های موقتی صورت بر اساس حرکت فرد در حین صحبت پیشنهاد شد [71]. یک روش جدید استخراج ویژگی برای طبقه‌بندی موثر تصاویر دیپ‌فیک پیشنهاد شده بود [72]. در پژوهش

---

[1] Mean Squared Error (MSE)
[2] Loss function
[3] Recurrent Neural Network (RNN)
[4] Optical flow based technique
[5] Patch and pair convolutional neural networks

[73]، یک رویکرد چندوجهی برای تشخیص ویدیوهای واقعی و دیپ‌فیک پیشنهاد شد. این روش شباهت‌های بین حالت‌های صوتی و تصویری را در یک ویدیو استخراج و تجزیه‌وتحلیل می‌کند. در تحقیق [74]، یک روش تشخیص محتوای دیپ‌فیک برای یافتن اختلافات بین چهره‌ها و زمینه آن‌ها با ترکیب چندین مدل *XeptionNet* اعمال شد. در [75] یک شبکه کانولوشن قابل تفکیک برای تشخیص چنین دستکاری‌هایی در تصاویر استفاده شده است. برای دسته‌بندی بهتر چهره‌های جعلی به تابع از ضرر سه‌گانه فرآیند استخراج ویژگی متوسل می‌شود [76]. در [77] طبقه‌بندی کننده مبتنی بر پچ (*patch-based classifier*) معرفی شد که به‌جای تمرکز روی ساختار کلی، بر روی پچ‌های محلی تصویر تمرکز می‌کند. در [78]، [79]، نویسندگان ویژگی‌ها را با استفاده از شبکه‌های *VGG* بهبود یافته استخراج کردند. یک آزمون نهایی برای تشخیص ویدئو دیپ‌فیک از طریق تشخیص ابروها نیز در [80] انجام شد.

۳-۱-۲- روش‌های مبتنی بر اندازه‌گیری‌های آماری

تعیین معیارهای آماری مختلف مانند میانگین نمرات همبستگی متقابل نرمال شده[1] بین داده‌های اصلی و مشکوک به اصلی، به تعیین اصالت و جعل بودن آن‌ها کمک می‌کند. کوپمن و همکاران [81] عدم یکنواختی پاسخ عکس[2] را برای تشخیص جعل در فریم‌های ویدیو بررسی کردند. *PRNU* یک الگوی نویز منحصربه‌فرد در تصاویر دیجیتالی است که به دلیل نقص در سنسورهای حساس به نور دوربین رخ می‌دهد. به دلیل متمایز بودن، آن را اثر انگشت عکس‌های دیجیتال به حساب می‌آوردند. این تحقیق دنباله‌ای از فریم‌ها را از ویدیوهای ورودی تولید می‌کند و آن‌ها را در فهرست‌های دسته‌بندی موقتی ذخیره می‌کند. هر فریم ویدئو با همان محدوده پیکسلی بریده می‌شود تا بخشی از دنباله *PRNU* حفظ شود. سپس این فریم‌ها به هشت گروه مساوی تقسیم می‌شوند. سپس الگوی استاندارد *PRNU* را برای هر فریم با استفاده از روش *FSTV* مرتبه دوم می‌سازد [82]. پس‌ازآن، با اندازه‌گیری نمرات همبستگی متقاطع نرمال شده و محاسبه تفاوت بین نمرات همبستگی و میانگین امتیاز همبستگی برای هر فریم، آن‌ها را محاسبه می‌کند. برای ارزیابی ویژه آماری بین دیپ‌فیک‌ها و ویدیوهای اصلی، نویسندگان بر روی نتایج یک *t-test* انجام می‌دهند [83]. برای مدل‌سازی یک ساختار پایه‌ای تولیدکننده‌ی کانولوشنال، نویسندگان در [84] مجموعه‌ای از ویژگی‌های ناحیه‌ای را با استفاده از الگوریتم انتظار-فرض ماکزیمم (*EM*) استخراج کردند. پس از استخراج، آن‌ها اعتبارسنجی اختصاصی را روی این معماری‌ها (مانند *GDWCT*، *STARGAN*، *ATTGAN*، *STYLEGAN* و *STYLEGAN2*) با استفاده از طبقه‌بندی‌کننده‌های ساده در آزمایش‌های اولیه اعمال می‌کنند. آگاروال و همکاران [85] با پیشنهاد چارچوب آماری [86] برای تشخیص دیپ‌فیک، یک آزمون فرضیه انجام دادند. ابتدا، این روش کوتاه‌ترین مسیر بین توزیع تصاویر اصلی و تصاویر ساخته‌شده با شبکه‌های مولد (*GAN*) *antagonistic* را تعریف می‌کند. بر اساس نتایج این فرضیه، این فاصله توانایی تشخیص را اندازه‌گیری می‌کند. به‌عنوان‌مثال، هنگامی‌که این فاصله افزایش یابد، می‌توان به راحتی دیپ‌فیک را تشخیص داد. معمولا اگر *GAN* مقدار کمتری از صحت را ارایه دهد، فاصله افزایش می‌یابد. علاوه بر این، یک *GAN* بسیار دقیق برای ایجاد تصاویر دستکاری شده با وضوح بالا که تشخیص آن‌ها سخت‌تر است الزامی است.

۴-۱-۲- روش‌های مبتنی بر بلاک چین

فناوری بلاک چین ویژگی‌های مختلفی را ارایه می‌کند که می‌تواند منشا محتوای دیجیتال را به شیوه‌ای بسیار قابل اعتماد، ایمن و غیرمتمرکز تایید کند. در فناوری بلاک چین عمومی، هر کسی به هر تراکنش، گزارش و رکورد دسترسی مستقیم دارد. برای تشخیص دیپ‌فیک، بلاک چین عمومی یکی از مناسب‌ترین راه‌حل‌های تکنولوژیکی برای تایید ویدیو یا اصالت تصویر است. به‌طورکلی کاربران معمولا باید منشا ویدیوها یا تصاویر را که به‌عنوان مشکوک علامت‌گذاری می‌شوند، کشف کنند.

حسن و صلاح [87] یک چارچوب عمومی مبتنی بر بلاک چین را برای ردیابی منشا ویدیوهای مشکوک به منابع خود پیشنهاد کردند. این راه‌حل پیشنهادی می‌تواند سوابق تراکنش خود را ردیابی کند، حتی اگر مطالب چندین بار کپی شده باشد. اصل اساسی می‌گوید که محتوای دیجیتال زمانی معتبر تلقی می‌شود که به‌طور قانع‌کننده‌ای در یک منبع قابل اعتماد ردیابی شود. برای دیپ‌فیک‌ها، بلاک چین عمومی اصالت محتوای ویدیویی را

[1] Average normalized cross-correlation scores between original and suspected data

[2] Photo Response Nonuniformity (PRNU)

به روشی غیرمتمرکز تایید می‌کند، زیرا این فناوری می‌تواند برخی از ویژگی‌های حیاتی را برای اثبات صحت آن ارایه دهد. در زیر ویژگی‌ها اصلی آمده است:

۱. یک چارچوب عمومی مبتنی بر فناوری بلاک چین را با تنظیم مدرکی مبنی بر اصالت محتوای دیجیتال در منبع مورد اعتماد آن ارایه می‌کند.
۲. جزییات معماری و طراحی راه‌حل پیشنهادی را برای کنترل و مدیریت تعاملات و تراکنش‌ها بین شرکت‌کنندگان ارایه می‌دهد.
۳. ویژگی‌های حیاتی قابلیت ذخیره‌سازی غیرمتمرکز مبتنی بر *IPFS* را با سرویس نام اتریوم مبتنی بر بلاک چین ادغام می‌کند [88].

چان و همکاران [89] یک رویکرد غیرمتمرکز مبتنی بر بلاک چین برای ردیابی منشا تاریخی محتوای دیجیتال (به‌عنوان‌مثال، تصویر، ویدئو و ...) پیشنهاد کردند. در این رویکرد پیشنهادی، چندین شبکه *LSTM*[1] به‌عنوان یک رمزگذار عمیق برای ایجاد ویژگی‌های متمایز استفاده می‌شوند که سپس فشرده شده و برای هش تراکنش استفاده می‌شوند. ویژگی‌های اصلی این مقاله به شرح زیر است:

۱. با استفاده از چندین مدل *LSTM CNN*، محتوای تصویر/ ویدیو هش و کدگذاری می‌شود.
۲. ویژگی‌های ابعادی بالا به‌عنوان یک ساختار کدگذاری شده باینری حفظ می‌شوند.
۳. اطلاعات در یک بلاک‌چین مبتنی بر دسترسی مجوز ذخیره می‌شود که به مالک کنترل محتوای آن را می‌دهد.

بر اساس مطالعات انجام شده، با جمع‌آوری تمامی این روش‌ها، **جدول ۱** دسته‌بندی‌های استراتژی‌های تشخیص دیپ‌فیک را فهرست می‌کند و تعداد و درصد (*PCT*) دسته‌بندی‌های مرتبط مطالعات را نشان می‌دهد. این جدول شامل ۸۸ مطالعه است، به استثنای برخی موارد که روش‌های مختلف را ادغام می‌کند که مشترک هستند. همچنین، این جدول نشان می‌دهد که رویکرد مبتنی بر یادگیری عمیق پرکاربردترین تکنیک است که حدود ۷۷% در تمام مطالعات را به خود اختصاص داده است. تحقیقات مربوط به رویکردهای یادگیری ماشین و روش‌های آماری به ترتیب ۱۸% و ۳% است. تعداد مطالعات در این تحلیل بر روی رویکرد مبتنی بر بلاک چین ۲% است. به‌طورکلی، ما تکنیک‌های تشخیص دیپ‌فیک را به چهار دسته تقسیم می‌کنیم که شامل روش‌های مبتنی بر یادگیری عمیق، تکنیک‌های مبتنی بر یادگیری ماشین، تکنیک‌های مبتنی بر آمار و تکنیک‌های مبتنی بر بلاک چین است. در میان آن‌ها، روش‌های مبتنی بر یادگیری عمیق به‌طور گسترده‌تری برای تشخیص دیپ‌فیک استفاده می‌شوند.

جدول ۱- دسته‌بندی روش‌های شناسایی دیپ‌فیک.
Table 1- Classification of deepfake identification methods.

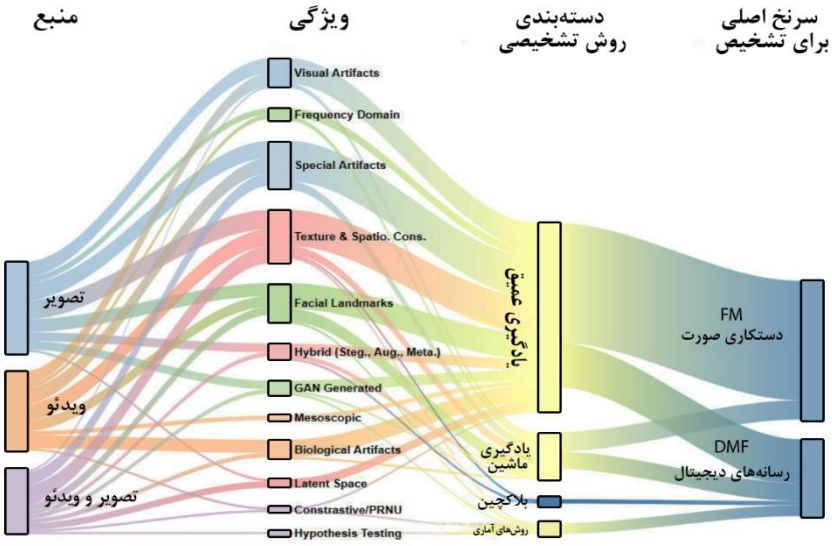
روش / مدل	رفرنس مطالعه	تعداد	درصد
روش‌های بر اساس یادگیری عمیق	[12]، [19]، [20]، [24]، [53]-[27]، [67]-[55]، [75]-[70]، [79]-[77]، [90]-[106]	71	%77
روش‌های بر اساس یادگیری ماشین	[23]-[17]، [91]، [92]، [113]-[107]	16	%18
روش‌های بر اساس آماری	[80]، [83]، [84]	3	%3
روش‌های بر اساس بلاک‌چین	[86]، [94]	2	%2

۲-۲- سوال: چارچوب طبقه‌بندی برای رویکردهای مختلف در تشخیص دیپ‌فیک چیست؟

برای درک بهتر، یافته‌های کلیدی خود را در **شکل ۲** خلاصه می‌کنیم. همان‌طور که در شکل نشان داده شده است، ما رویکردهای کلی مربوط به عناصر مختلف مانند داده‌های ورودی، ویژگی‌ها، روش و نوع تکنیک‌ها را طبقه‌بندی می‌کنیم. یک مسیر بین دو عنصر نشان‌دهنده اجزای مرتبط مورد استفاده در مقاله همراه برای هر روش است. همان‌طور که در شکل نشان داده شده است، بیشتر مقالات تصویر یا ویدئو را به‌عنوان داده ورودی اعمال می‌کنند، درحالی‌که بسیاری از مقالات از تصویر و ویدئو به‌عنوان ورودی استفاده می‌کنند. حدود ۷۵% از روش‌ها از تکنیک‌های مبتنی بر یادگیری عمیق به‌عنوان روش تشخیص استفاده می‌کردند. تنها چند مقاله از روش‌های بلاک چین و آماری برای شناسایی و تشخیص دیپ‌فیک‌ها استفاده

[1] Long short-term memory

کردند. در تشخیص دیپ‌فیک، تکنیک‌های زیربنایی مختلفی در دسترس هستند، مانند سیگنال‌های بیولوژیکی، عدم تطابق *Phoneme-Viseme*[1]، حالت‌ها و حرکات صورت (به‌عنوان‌مثال، موقعیت‌های نقطه عطف صورت دوبعدی و سه‌بعدی، حالت سر و واحدهای عمل صورت) بااین‌حال، روش‌های مبتنی بر یادگیری ماشین تقریبا به‌طور مساوی از هر دو تکنیک استفاده می‌کنند. مشترک برای هر دو رویکرد بلاک‌چین و آماری، آن‌ها فقط از رسانه‌های دیجیتال به‌عنوان بخشی از تکنیک تشخیص استفاده می‌کنند.



شکل ۲- طبقه‌بندی تکنیک‌های تشخیص دیپ‌فیک.

Figure 2- Classification of deepfake detection techniques.

این طبقه‌بندی الگوریتم‌های تشخیص را بر اساس رسانه (تصویر، ویدئو، یا تصویر و ویدئو)، ویژگی‌های مورد استفاده (از بین ۱۲ ویژگی)، روش تشخیص (*DL*، *ML*، بلاک چین یا آماری) و سرنخ برای تشخیص (دستکاری چهره و رسانه‌های دیجیتال، یا نشانه‌های دیگر) دسته‌بندی می‌کند. اندازه خط اتصال تعداد نسبی مقالات و پژوهش‌ها را منعکس می‌کند.

جمع‌بندی و نگاه اجمالی به **شکل ۲** به شرح زیر است:

۱. رسانه‌های دیجیتال و دستکاری صورت، سرنخ اصلی برای تشخیص را نشان می‌دهند (*DMF* یا *Digital Media Forensics* تشخیص جعل رسانه دیجیتال و *FM* یا *Facial Manipulation* دستکاری چهره است).

۲. ستون دسته‌بندی روش تشخیصی که دسته روش و مدل‌ها را نشان می‌دهد (*ML*: یادگیری ماشین، *DL*: یادگیری عمیق، *STAT*: روش آماری، *BC*: بلاک‌چین) که شامل:

- *DL*: روش و الگوریتم‌های یادگیری عمیق؛ مانند CNN: شبکه عصبی کانولوشنال، *RNN*: شبکه عصبی بازگشتی، *RCNN*: شبکه عصبی کانولوشنال ناحیه‌ای، *MTCNN*: شبکه‌های عصبی کانولوشنال چندوظیفه‌ای آبشاری، *MSCNN*: شبکه عصبی کانولوشنال زمانی چند مقیاس)
- *ML*: روش و الگوریتم‌های یادگیری ماشین؛ مانند *SVM*: ماشین بردار پشتیبان، *RF*: جنگل تصادفی، *MLP*: شبکه عصبی پرسپترون چندلایه، *LR*: رگرسیون لجستیک، *k-MN*: خوشه‌بندی *K*-میانگین، *XGB: XGBoost*، *ADB: AdaBoost*، *DT*: درخت تصمیم‌گیری، *NB*: بیز ساده، *KNN*: همسایه نزدیک *K*، *DA*: تحلیل تشخیص.
- *STAT*: روش‌ها و الگوریتم‌های آماری؛ مانند *EM*: امیدبخشی-فرض ماکزیمم، *CRA*: تحلیل همبستگی.
- *BC*: روش و الگوریتم‌های مبتنی بر بلاک‌چین؛ مانند *ETH*: بلاک‌چین اتریوم.

۳. ویژگی‌ها: (*VA*: مصنوعات بصری، *FD*: تحلیل حوزه فرکانس، *SA*: مصنوعات خاص، *STC*: سازگاری فضایی-زمانی، *FL*: نقاط شاخص چهره، *TEX*: بافت، *FDA*: تحلیل حوزه فرکانس، *LS*: ویژگی نهفته، *GAN*: ویژگی مبتنی بر شبکه‌های مولد متخاصم، *MES*: ویژگی‌های مزوسکوپی، *BA*: مصنوعات

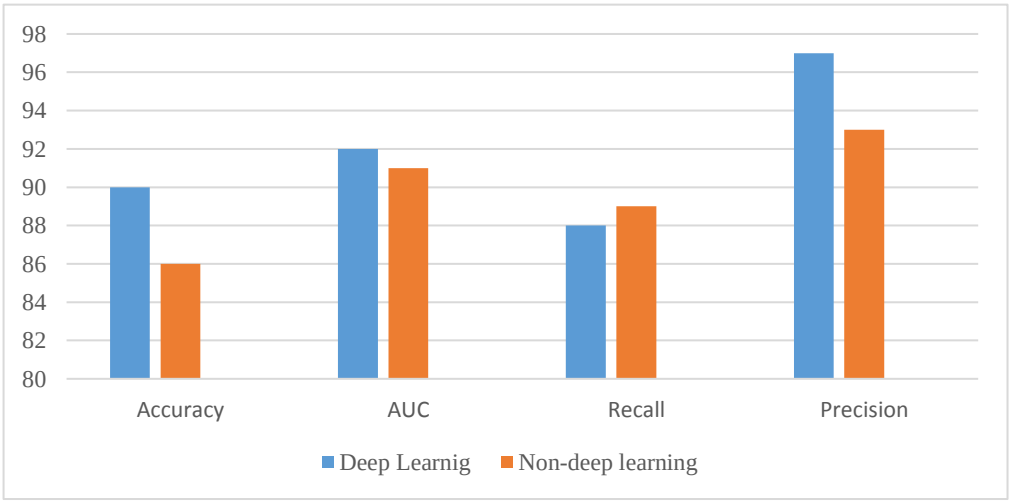
[1] Phoneme-Viseme Mismatches

بیولوژیکی، *HT*: آزمایشات بر روی فرضیه، *IFIC*: ناسازگاری درون فریم، *CPRNU*: الگوی *PRNU* متضاد و حساس به نور، *IMG*: فرا داده تصویر، افزایش و *Steganalysis*، موارد دیگر: ویژگی‌های مختلف خارج از لیست مشترک).

۴. مجموعه داده‌ها: (*FF: FaceForensics*، *FFCC: FaceForensicsCC*، *DFD*: تشخیص دیپ‌فیک، *CELEB-A*: تشخیص جعل عمیق نسخه ۱، *CELEB-DF*: تشخیص جعل عمیق نسخه ۲، *DFDC*: چالش تشخیص دیپ‌فیک، *DF-TIMIT: Deepfake-TIMIT*، *DF-1.0: DeeperForensics-1.0*، *WDF: Wild Deepfake* یا *SMFW*، *SwapMe* و *FaceSwap*، *DFS*: جعل‌های عمیق، *FFD*: چهره‌های جعلی در طبیعت، *FE: FakeET*، *FS* یا *Face Shifter*، *DF:* Deepfake، *SFD*: تشخیص چهره تعویض شده، *UADFV*: حالت‌های ناسازگار، *MANFA*: چهره دستکاری شده) و مواردی از این دست را پوشش می‌دهد.

۳-۲- سوال: بر اساس نتایج به‌دست‌آمده در مقالات و مدل‌های مختلف، آیا کارایی مدل‌های مبتنی یادگیری عمیق در تشخیص دیپ‌فیک بهتر از مدل‌های غیراز این است؟

در این پرسش مدل‌ها به دو گروه تقسیم شده مدل‌های مبتنی بر یادگیری عمیق مدل‌های مبتنی بر یادگیری غیرعمیق. میانگین دقت *Accuracy*، مساحت زیر منحنی *AUC*، بازیابی *Recall* و ارزیابی *Precision* را محاسبه کردیم. سپس، یک تحلیل تطبیقی روی عملکرد این دو گروه مدل اجرا کرده و نتایج میانگین را به دست آوردیم. بر اساس ارزیابی این مدل‌ها با استفاده از معیارهای عملکرد (دقت، حساسیت و ارزش درستی) مشاهده کردیم که به‌طورکلی مدل‌های مبتنی بر یادگیری عمیق عملکرد بهتری نسبت به مدل‌های مبتنی بر یادگیری غیر عمیق دارند. همان‌طور که نتایج در **شکل ۳** نشان داده شده است، عملکرد دقت و ارزش درستی در مدل‌های یادگیری عمیق به‌طور قابل توجهی بهتر از مدل‌های یادگیری غیر عمیق است. بااین‌حال، در مورد *AUC* و بازیابی، عملکرد آن‌ها تقریبا مشابه است. نتایج کلی برتری مدل‌های مبتنی بر یادگیری عمیق نسبت به مدل‌های مبتنی بر یادگیری غیر عمیق را نشان می‌دهد.



شکل ۳- مقایسه نتایج بین مدل‌های بر اساس یادگیری عمیق و (غیر از آن) یادگیری غیرعمیق.

Figure 3 - Comparison of results between models based on deep learning and (other than) non-deep learning.

۴-۲- سوال: برای تشخیص دیپ‌فیک از چه مدل‌هایی استفاده می‌شود؟

این بخش به شرح مدل‌های مختلفی می‌پردازد که برای تشخیص دیپ‌فیک استفاده می‌شوند. بر اساس این مطالعه، ما این مدل‌ها را به سه گروه اصلی تقسیم می‌کنیم شامل مدل یادگیری عمیق، مدل یادگیری ماشین و مدل آماری است.

مدل‌های یادگیری عمیق

در بینایی ماشین[1]، مدل‌های یادگیری عمیق به دلیل مکانیزم استخراج و انتخاب ویژگی، به‌طور گسترده‌ای مورد استفاده قرار گرفته‌اند، زیرا آن‌ها می‌توانند مستقیما ویژگی‌ها را از داده استخراج کنند یا اینکه یاد بگیرند. در مطالعات تشخیص دیپ‌فیک، مدل‌های یادگیری عمیق زیر مورد استفاده قرار گرفته‌اند:

۱. مدل شبکه عصبی کانولوشنال (*CNN*): به‌عنوان‌مثال، *XceptionNet*، *GoogleNet*، *VGG*، *ResNet*، *EfficienNet*، *HRNet*، *InceptionResNetV2*، *MobileNet*، *InceptionV3*، *DenseNet*، *SuppressNet*، *StatsNet*.
۲. مدل شبکه عصبی بازگشتی (*RNN*): به‌عنوان‌مثال *LSTM*، *FaceNet*.
۳. مدل *RNN* دوطرفه یا *Bidirectional RNN*.
۴. مدل شبکه عصبی بازگشتی کانولوشنال بلندمدت یا (*RCNN*) *Long-term Recurrent Convolutional Neural Network*.
۵. مدل *Faster RCNN*.
۶. مدل شبکه حافظه سلسله مراتبی ((*HMN*) *Hierarchical Memory Network*).
۷. مدل شبکه‌های عصبی کانولوشنال چندوظیفه‌ای آبشاری ((*MTCNN*) *Multi-task Cascaded CNNs*).
۸. یادگیری عمیق مجموعه‌ای ((*DEL*) *Deep Ensemble Learning*).

مدل یادگیری ماشین

این تکنیک با استفاده از الگوریتم‌های پیشرفته انتخاب ویژگی، بردار ویژگی را با تعریف ویژگی‌های درست ایجاد می‌کند. سپس این بردار را به‌عنوان ورودی برای آموزش طبقه‌بندی‌کننده تغذیه می‌کند تا ویدیوها یا تصاویر دستکاری شده توسط دیپ‌فیک را از موارد واقعی تشخیص دهد. موارد زیر به‌عنوان مدل‌های یادگیری ماشین استفاده می‌شوند:

۱. ماشین بردار پشتیبان (*SVM*).
۲. رگرسیون لجستیک (*LR*).
۳. پرسپترون چندلایه (*MLP*).
۴. تقویت تطبیقی (*AdaBoost*).
۵. تقویت گرادیان شدید (*XGBoost*).
۶. خوشه‌بندی (*k-MN*) *K-Means*.
۷. جنگل تصادفی (*RF*).
۸. درخت تصمیم‌گیری (*DT*).
۹. تحلیل تشخیصی (*DA*).
۱۰. بیز ساده (*NB*).
۱۱. یادگیری چند نمونه‌ای (*MIL*).

مدل آماری

مدل‌های آماری بر پایه استفاده از مطالعات تئوری اطلاعات برای اعتبارسنجی بنا شده‌اند. در این مدل‌ها، کوتاه‌ترین مسیر بین توزیع داده‌های ویدیو/تصویر اصلی و دستکاری‌شده با دیپ‌فیک محاسبه می‌شود. به‌عنوان‌مثال، در [81] برای میانگین نمره‌های همبستگی متقاطع نرمال‌شده بین ویدیوی اصلی و دیپ‌فیک، یک مقدار اهمیت اندازه‌گیری می‌شود تا آن‌ها را به‌عنوان جعلی یا واقعی طبقه‌بندی کند. برخی از مدل‌های آماری پرکاربرد عبارتند از:

۱. امیدبخشی-فرض ماکزیمم ((*EM*) *Expectation-Maximization*).

[1] Computer vision

۲. فاصله کل اختلاف (*Total Variational (TV) distance*).

۳. واگرایی کولباک-لیبلر (*Kullback-Leibler (KL) divergence*).

۴. واگرایی جنسن-شانون (*Jensen-Shannon (JS) divergence*).

بر اساس این مطالعات، دسته‌بندی‌هایی را در مدل‌های یادگیری عمیق، مدل‌های یادگیری ماشین و روش‌های آماری انجام می‌دهیم، همان‌طور که در **جدول ۲** نشان داده شده است. این جدول تعداد و درصد مدل‌های استفاده شده در مطالعات را به غیر از ۲۱ بررسی متفاوت، مشخص می‌کند. همچنین مشاهده می‌کنیم که مطالعات مبتنی بر یادگیری عمیق (*DL*) بالاترین نسبت را در این پژوهش‌ها دارند.

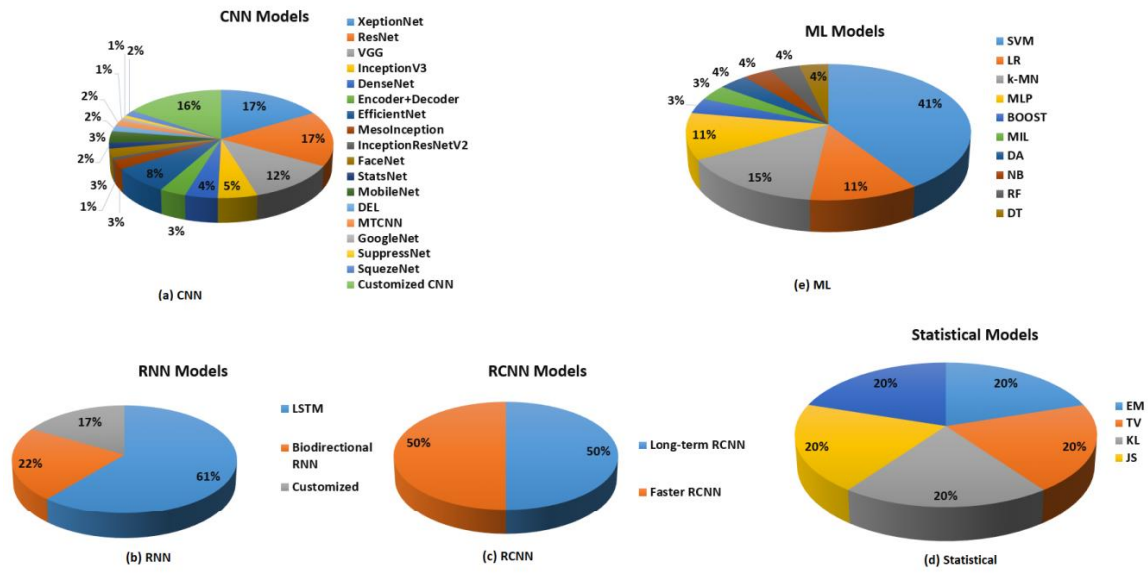
جدول ۲- توزیع مدل‌های استفاده شده.

Table 2- Distribution of used models.

دسته‌بندی	مدل	مورد مطالعه	درصد
یادگیری عمیق	CNN	71	%78
	RNN	12	%13
	RCNN	2	%2
یادگیری ماشین	SVM	11	%12
	K-MN	4	%4
	LR	3	%3
	MLP	3	%3
	BOOST	2	%2
	RF	1	%1
	DT	1	%1
	DA	1	%1
	NB	1	%1
	MIL	1	%1
آماری	EM	1	%1
	TV, KL, JS	1	%1

شکل ۴ نسخه‌های کامل گروه‌های آشکارساز (شناسایی کننده) دیپ‌فیک را که از این مطالعات به‌دست‌آمده‌اند نشان می‌دهد، جایی که *CNN* بیشترین زیرمجموعه را دارد. بر اساس **شکل ۴**، ما زیرمجموعه‌بندی بیشتری را روی مدل‌های *CNN* انجام می‌دهیم و سه مدل *CNN* زیر را پیدا می‌کنیم: *XceptionNet* و *ResNet* به ترتیب هرکدام ۱۷% و *VGG* با ۱۲% هستند. علاوه بر این، مدل‌های *LSTM* ، ۱۳% از *RNN* را تشکیل می‌دهند. علاوه بر این، محبوب‌ترین مدل یادگیری ماشین *SVM* با ۱۲% و *k-MN* با ۴% است. توزیع جزییات در مدل‌های مختلف در **شکل ۴** ارایه شده است که نسبت مدل‌های استفاده‌شده (مانند *DL*، *ML* و آماری) را در مطالعات مختلف برای تشخیص دیپ‌فیک نشان می‌دهد. علاوه بر این، پاسخی برای *SRQ-2.3* ارایه می‌دهد. مقالات بررسی‌شده نشان می‌دهند که مدل‌های شبکه عصبی عمیق (*DNN*) در تشخیص دیپ‌فیک موفق هستند، جایی که مدل‌های مبتنی بر *CNN* در بین تمام مدل‌های *DNN* کارایی بیشتری را نشان می‌دهند.

در **شکل ۵**، که یک نمودار درختی از لیست مدل‌ها و الگوریتم‌های تشخیص دیپ‌فیک است، به‌طور مشخص سه روش اصلی برای تشخیص دیپ‌فیک به همراه الگوریتم‌های آن‌ها آمده است. **شکل ۵** نسخه کامل گروه‌های آشکارساز را نشان می‌دهد که از این مطالعات اولیه پیدا شده‌اند، جایی که *CNN* بیشترین تقسیم‌بندی را دارد. بر اساس **جدول ۲**، ما یک طبقه‌بندی فرعی بر روی مدل‌های *CNN* اعمال کردیم و دریافتیم که ۳ مدل *CNN* مقادیر شامل *XeptionNet* (*i*) و *ResNet* به ترتیب ۱۷% و *VGG* با ۱۲% را دارند. علاوه بر این، مدل‌های *LSTM*، ۱۳% از *RNN* را می‌گیرند. علاوه بر این، محبوب‌ترین مدل یادگیری ماشین *SVM* با ۱۲% و *k-MN* با ۴% است. توزیع جزییات در مدل‌های مختلف در **شکل ۴** ارایه شده است که نسبت مدل‌های استفاده شده (به‌عنوان‌مثال، *DL*، *ML* و آماری) را در مطالعات مختلف برای تشخیص دیپ‌فیک نشان می‌دهد. مقالات بررسی شده نشان می‌دهند که مدل‌های شبکه عصبی عمیق (*DNN*) در تشخیص دیپ‌فیک موفق هستند، جایی که مدل‌های مبتنی بر *CNN* کارایی بیشتری را در بین تمام مدل‌های *DNN* نشان می‌دهند.



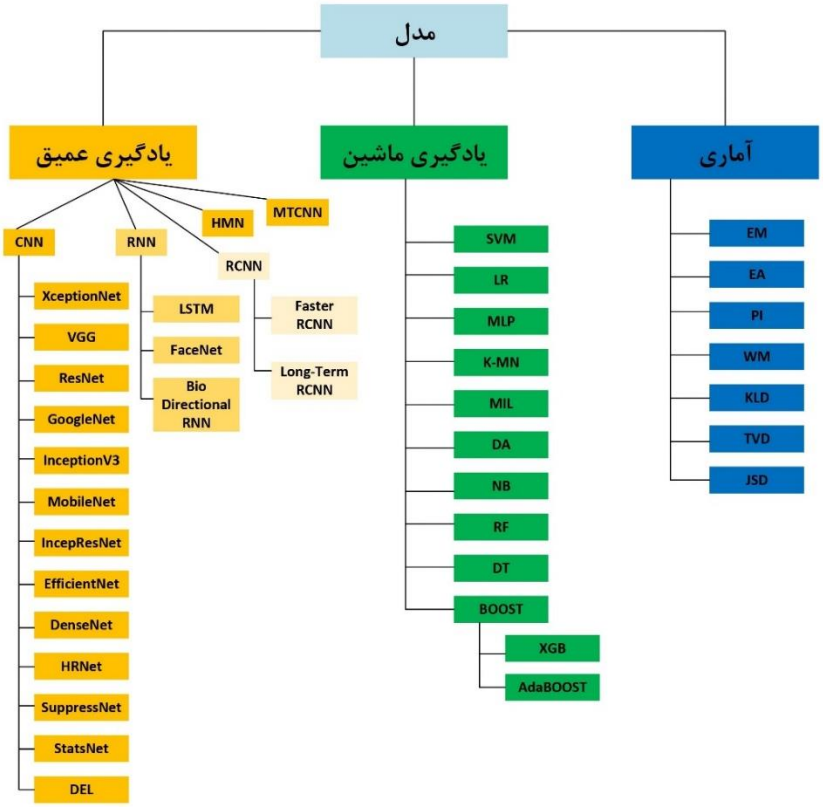
شکل ۴- تخصیص زیرمجموعه‌های مدل‌های تشخیص: ML: یادگیری ماشین و DL: یادگیری عمیق.

Figure 4- Allocation of subsets of recognition models: ML: machine learning and DL: deep learning.

۳- روش‌شناسی تحلیل رسانه‌ای دیپ‌فیک

هدف این بخش از پژوهش حاضر درک بهتر تلاقی رسانه‌های دیجیتال و هوش مصنوعی و پیامدهای آن در جامعه مدرن است. به این منظور می‌توان از روش‌های تحقیقی گوناگون استفاده کرد [95]. برای تجزیه‌وتحلیل مستنداتی انتخاب شده که دسترسی به آن امکان‌پذیر است. یکی از چالش‌ها این است که با توجه به اینکه پدیده دیپ فیک بسیار جدید است به همین دلیل، ادبیات بررسی شده محدود است. در ادبیات اصلی به صراحت پرداختن به دیپ‌فیک موردتوجه قرار گرفته است، به‌عنوان‌مثال، [4]–[1]، [96]، [97] که با تجزیه‌وتحلیل چندین سند دیگر از جمله گزارش‌های فنی، بیانیه‌های مطبوعاتی، پیشنهادات قانونی، سمعی و بصری هستند تکمیل شد. همچنین مطالب موجود در پلتفرم‌های اینترنتی با دسترسی آزاد مانند یوتیوب نیز مورد استفاده قرار گرفته است.

همان‌طور که پیش‌تر بیان شد هدف این است که یافته‌ها، دیدگاه‌ها، مثال‌ها را از دیدگاه‌های مختلف بررسی و ترکیب کنیم، علاوه بر این، درمجموع به نقد سازمان‌دهی شده‌ای در رابطه با این پدیده بپردازیم. در ۷ بخش پیش رو از دیدگاه‌های مختلف این موضوع بررسی شده است. برای هر دیدگاه، ادبیات اضافی موردتوجه و بررسی قرار می‌گیرد که ویژگی‌های کلیدی مرتبط با اصل آن دیدگاه را به تصویر می‌کشد. زوایای مختلف انتخاب شده برای بررسی این پدیده، الهام گرفته از رویکرد تقاطع پذیری هستند، رویکردی که می‌تواند برای بررسی چگونگی ارتباط بین مرتبط بودن دسته‌ها و نحوه تعامل آن‌ها در سطوح مختلف مورد استفاده قرار گیرد. از طریق تجزیه‌وتحلیل متقاطع، می‌توان تلاش کرد تاثیر حوزه‌های مختلف رسانه‌های دیجیتال و پدیده‌های اجتماعی دخیل مانند بی‌عدالتی، نابرابری، نفوذ سیاسی و غیره را شناسایی و درک کرد. برای بررسی یک پدیده جدید و پیچیده مانند دیپ فیک، به دیدگاه‌های متعددی نیاز است تا بتوان ماهیت آن را به دست آورد. بنابراین، تصمیم بر این شد که از دیدگاه‌هایی که رسانه‌های دیجیتال می‌توانند ارایه دهند، یعنی رسانه و جامعه، تولید رسانه، نمایندگی رسانه و مخاطبان رسانه، شروع کنیم. بااین‌حال، ازآنجایی‌که جنبه‌های اجتماعی در کانون توجه قرار دارند، دیدگاه‌های اضافی مانند جنسیت، سیاست و همچنین قانون و مقررات ضروری تلقی می‌شود. درحالی‌که این‌ها ممکن است جامع نباشد، ما فکر می‌کنیم که مبنای کافی برای بحث‌هایی فراهم می‌کنند که تا حدودی این پدیده را روشن می‌کند و ما را قادر می‌سازد در مورد پیامدهای آن بحث کنیم.



شکل ۵- لیست مدل‌ها و الگوریتم‌های تشخیص دیپ‌فیک.

Figure 5- List of deepfake detection models and algorithms.

البته باید به این نکته اشاره کرد که رویکرد بررسی ما در این حوزه بیشتر به‌صورت آزادانه دنبال می‌شود تا به‌طور دقیق اجرا شود (اگرچه از تقاطع مطالب نیز الهام گرفته شده) و این به هدف ارایه یک دید چندوجهی و تحلیل یک پدیده روزمره از دیدگاه‌های مختلف کمک می‌کند. شناسایی این پیامدها اهمیت دارد، زیرا از نظر روش‌شناسی، جداسازی و شناسایی اثرات رسانه‌ها در سراسر فضای اجتماعی دشوار یا زمانبر است [98].

۱-۳- دیدگاه رسانه و جامعه

فرهنگ دایما تغییر می‌کند و تاثیر آن در جامعه با نحوه‌ی تعامل مردم و استفاده از فناوری‌های مدرن درهم‌تنیده است. ثبت فرآیند تغییرات در فرهنگ عامه تحت تاثیر رسانه‌ها که اغلب به‌صورت بلندمدت اتفاق می‌افتد کار دشوار و مهمی است [99]. رسانه‌های جمعی از روزنامه‌ها و تلو یزیون در گذشته به شبکه‌ها و پلتفرم‌های دیجیتالی تبدیل شده‌اند که به‌عنوان یک فضای مشترک دیداری و ارتباط افزوده تلقی می‌شوند [100]. به‌طورکلی، رسانه‌های جمعی دارای اثراتی هستند که منجر به تغییر در یک نتیجه در یک شخص یا نهاد اجتماعی می‌شود که به دلیل نفوذ و تاثیر رسانه‌های جمعی و در معرض چشم قرارگرفتن پیام‌های آن‌ها در جامعه است [101]. به‌این‌ترتیب، محتوایی که از طریق هر نوع رسانه جمعی منتقل می‌شود، مثلا دیپ‌فیک‌ها در رسانه‌های اجتماعی، نگرانی‌های اساسی ایجاد می‌کند. از دیدگاه رسانه‌ها، تلاش‌ها برای به تصو یر کشیدن تاثیر هوش مصنوعی بر زندگی افراد یک جامعه، به‌ویژه از دیدگاه اجتماعی، سیاسی یا اخلاقی، قبلا در داستان‌های علمی تخیلی و رمان‌ها مطرح بود. به این صورت که هوش مصنوعی اغلب در وجود ماشین‌های هوشمند تجسم می‌یافت که به‌طور مستقل وظایف بی‌اهمیت را مدیریت می‌کنند [102].

بر این اساس فرآیندهای تکنولوژیک و موثر بر مسایل اجتماعی و تنظیم آن‌ها باید در منظری جدید موردتوجه قرار گیرد. امروزه، مقررات محدودی در قبال استفاده از دیپ فیک در کشورهای جهان وجود دارد که در فرآیندهای اجتماعی، آن‌ها در دسته بزرگتر میم‌های سرگرم کننده اینترنتی یا اخبار جعلی قرار می‌گیرند. از منظر بازی[1] نیز لازم به ذکر است که ما اکنون وارد عصر جدیدی می‌شو یم، جایی که این تعامل بسیار واقعی‌تر می‌شود، زیرا

[1] Game

سیستم‌های هوش مصنوعی بازی‌های پیچیده‌ای را علیه انسان‌ها انجام می‌دهند و برنده می‌شوند (به‌عنوان‌مثال، بازی رومیزی *Go*، بازی استراتژی *StarCraft II*). عملکرد انسانی در انجام وظایف خاص (به‌عنوان‌مثال، طبقه‌بندی تصویر) نیز این نشان می‌دهد که به زودی آنچه تاکنون به‌عنوان داستان و خیال‌پردازی از آینده تلقی می‌شد، به قلمرو دنیای واقعی قدم می‌گذارد، جایی که بر فرهنگ اجتماعی و تصمیم‌گیری‌های اجتماعی موجود نیز تاثیر دارد.

در بعد اقتصادی، اولین تاثیرات هوش مصنوعی به‌طور عام نیز در حال حاضر قابل مشاهده است، برخی فرآیندهای اقتصادی و تجاری در شرکت‌ها از طریق هوش مصنوعی بهبود می‌یابند و محصولات پیشرفته‌تری را ارایه می‌دهند، زیرا شرکت‌ها می‌توانند سودآوری و ریسک‌ها را به‌طور موثرتری مدیریت کنند [103]. در رسانه‌های دیجیتال مدرن، می‌توان مقادیر زیادی از اطلاعات دیجیتال را جمع‌آوری و تجزیه‌وتحلیل کرد، همچنین فرآیندها را می‌توان خودکار و حتی برای افراد سفارشی‌سازی کرد. به‌عنوان‌مثال، بسیاری از پلتفرم‌های شبکه‌های اجتماعی و سایر ارایه‌دهندگان خدمات، کمپین‌های بازاریابی هدفمند کاربران را بر اساس تاریخچه، اقدامات و اولویت‌های جست‌وجو یا سفارش‌های خرید کاربر مشخص می‌کنند که درواقع چنین پیام‌های سفارشی و انتخاب شده توسط تحلیل هوش مصنوعی می‌باشد. در بعد سیاسی، همچنین بسیاری از پیام‌های سیاسی سیاستمداران از دیدگاه‌های سایر شهروندان در رسانه‌های اجتماعی که توسط الگوریتم‌های پیشرفته هوش مصنوعی تحلیل می‌شود سرچشمه می‌گیرد.

تاثیر مستقیم دیپ فیک در اولین لایه‌ای که باید در نظر گرفته شود بی‌اعتمادی به سازمان‌ها، فرآیندها و افراد است که القا می‌شود. نگاه کردن به هوش مصنوعی از منظر رسانه‌ای و اجتماعی جالب است، به‌ویژه با توجه به اینکه امروزه بسیاری از موقعیت‌های مربوط به دیپ فیک به دلیل عدم تخصص و مشکل در مدیریت جنبه‌های پیچیده اجتماعی به اندازه کافی موردتوجه قرار نمی‌گیرند. هوش مصنوعی می‌تواند ما را قادر سازد تا افراد و دیدگاه‌های گروهی را بهتر درک کنیم، دانش آن‌ها را به موقع ترکیب کنیم، واکنش‌ها را شبیه‌سازی کنیم، و برای موقعیت‌های پیچیده تصمیم‌گیری درست داشته باشیم [104].

در تعامل بین هوش مصنوعی و جامعه همه‌چیز در یک سمت‌وسوی مثبت قرار نمی‌گیرد. بسیاری از مواقع فناوری‌های جدیدی که به‌خوبی درک نشده‌اند، مورد سوء استفاده قرار گرفته‌اند، زیرا فقدان چارچوب‌های نظارتی مناسب و قوانین بازدارنده یا جریمه کننده وجود نداشته‌اند، به‌عنوان‌مثال، برای روباتیک [105]. به خصوص در مطبوعات و همچنین در رسانه‌های اجتماعی، اغلب مقالاتی منتشر شده است که آینده‌ای دیستوپیایی، مبهم ناگوار و غیرقابل پیش‌بینی را ترسیم می‌کند، جایی که هوش مصنوعی همه‌چیز را کنترل می‌کند و اختیار را در انسان سرکوب می‌کند.

۲-۳- دیدگاه تولید رسانه

یکی از حوزه‌های کلیدی تحت تاثیر هوش مصنوعی در رسانه‌های دیجیتال، تولیدات رسانه است، جایی که یک رابطه دو طرفه بین تولید و مصرف وجود دارد. این سوال کلیدی که چه کسی تصمیم می‌گیرد چه چیزی تولید شود، اغلب از طریق فرآیندهایی که سعی در درک مخاطبان و ارایه محتوایی مطابق با علایق آن‌ها داشت، مطرح می‌شود. بااین‌حال، در عصر رسانه‌های دیجیتال، چنین فرآیندهایی خودکار هستند و به‌طور فوری با درنظر گرفتن تمام جزییات است. الگوریتم‌های هوش مصنوعی به‌صورت بلادرنگ حجم عظیمی از داده‌ها را تجزیه‌وتحلیل می‌کنند و پروفایل‌های دقیقی را در مورد کاربران، علایق، نیازها و رضایت آن‌ها استخراج می‌کنند [106].

با توجه به استخراج این پروفایل علایق، تولید محتواها نیز برنامه‌ریزی می‌شوند. به‌این‌ترتیب، دیپ‌فیک‌ها این هدف‌گیری را بیشتر تقویت می‌کنند، زیرا محتوا را می‌توان به‌گونه‌ای شکل داد که برای ویژگی‌های خاص یک فرد جذاب‌تر باشد. به‌عنوان‌مثال، پیامی می‌تواند در حمایت از یک نامزد سیاسی تولید شود که شامل یک بازیگر یا رهبر خاص است که کاربر آن را قابل اعتماد می‌داند و برایش قابل باور است. امروزه، سازمان‌های تجاری برای فرآیندهای تصمیم‌گیری خود به رسانه‌های اجتماعی تکیه می‌کنند و زمانی که با توانایی‌های یادگیری هوش مصنوعی همراه شود، رویکردهای هوش رسانه‌ای جدید می‌توانند ظاهر شوند که اجتماعی و چندوجهی هستند، بدین معنی که وجوه مختلف دستیابی به هدف را درنظر می‌گیرند که درنتیجه موثرتر خواهند بود [107].

خبر یک محصول است و نحوه تولید آن را می‌توان با نگاهی به قدرت در روابط فرهنگی، اقتصادی، سیاسی-اقتصادی یا سایر روابط درک کرد [108]. تاثیر متقابل رسانه و روزنامه‌نگاری بسیار پیچیده است و درک کامل آن ممکن است آسان نباشد. درک آنچه در رسانه‌های اجتماعی می‌گذرد و به چشم و گوش مخاطبان می‌رسد برای عموم مردم که در بسترهای رسانه‌های اجتماعی اخبار را دنبال می‌کنند آسان نیست و تاثیرات و پیامدهای آن به‌خوبی درک نشده است، همان‌طور که اخبار جعلی در رسانه‌هایی مثل توییتر[1] (ایکس) و فیس‌بوک[2] و سایر شبکه‌های اجتماعی در سال‌های اخیر در رابطه با انتخابات ۲۰۱۶ ایالات‌متحده این موضوع را نشان می‌دهد [109].

یکی دیگر از جنبه‌های کلیدی در تولیدات رسانه‌ای است که توسط هوش مصنوعی تولید یا تقویت می‌شود و توانایی ایجاد متن، عکس و ویدئو بر اساس آنچه از منابع موجود در پلتفرم‌های دیجیتال یاد می‌گیرد را دارد. به‌عنوان‌مثال، استخراج منابع از رسانه‌های اجتماعی مانند فیس‌بوک، توییتر، اینستاگرام و ... که با یادگیری عمیق و پیشرفت قابل توجه در پردازش زبان طبیعی (*NLP*) از جمله قابلیت خلاصه کردن متون [110] و همچنین بازتولید صداها بر اساس نمونه‌های محدود موجود (مثلا یک نمونه گفتار ۵ ثانیه‌ای) [111] که از طریق همین ۵ ثانیه می‌توان دیپ فیک ایجاد کرد، در سطحی که انسان‌ها آن را قانع کننده می‌دانند [112]. به‌این‌ترتیب، در عصری که اطلاعات دیجیتال به راحتی قابل دسترسی است و قابل کپی کردن است، ایجاد اخبار خودکار (از جمله زبان گفتاری) امکان‌پذیر است.

موارد این‌چنینی تاثیرات قابل توجهی بر تولیدات رسانه دارد، زیرا اکنون می‌توان با توجه به منابع کافی، به‌عنوان‌مثال، یک نمونه صدای در دسترس عموم از یک سخنرانی و یک عکس از رسانه‌های اجتماعی محتوای پیچیده توسط هوش مصنوعی ایجاد کرد؛ همچنین این می‌تواند منجر به ایجاد یک ویدیوی دیپ فیک شود [113]. در یک نظرسنجی [114]، برخی احساس منفی نسبت به اینکه ربات‌ها یا هوش مصنوعی به‌جایگاه روزنامه‌نگاری یا خبرنگاری آسیب خواهند زد ذهنیت منفی دارند، درحالی‌که برخی دیگر جنبه‌های مثبت آن را نیز می‌بینند. چنین چیزی برخی چالش‌ها را به همراه خواهد داشت، از جمله پرسش‌هایی مانند اینکه جهت‌گیری‌ها، ملاحظات اخلاقی و مسایلی از این دست چگونه توسط هوش مصنوعی مدیریت خواهد شد، زیرا هنوز درک روشنی از تعامل در بسیاری از الگوریتم‌های هوش مصنوعی و تصمیم‌های اتخاذ شده توسط آن‌ها وجود ندارد [115].

ازین رو امکان تولید محتوای مبتنی بر دیپ فیک افزایش پیدا می‌کند، درک مسایل اخلاقی و صحت سنجی برخی اخبار توسط انسان امکان‌پذیر است. پس خبرنگاری رباتی (متکی بر هوش مصنوعی) دارای پیامدهای عملی، اجتماعی سیاسی، روان‌شناختی، قانونی و شغلی قابل‌توجهی برای سازمان‌های خبری، روزنامه‌نگاران و مخاطبان آن‌ها است [116]. تحلیل دیگر این است که امروزه نیاز به کارمندانی وجود دارد که دارای مهارت‌های تولید رسانه‌های چندوجهی باشند [117]، اما اگر بتوان چنین وظایفی را خودکار کرد و (حتی تا حدی) به فرآیندهای مبتنی بر هوش مصنوعی محول کرد، نیاز به چنین مهارت‌هایی ممکن است ضروری نباشد یا باید تغییر کند، به‌عنوان‌مثال، تمرکز بر روی همکاری با ربات‌های هوش مصنوعی در تولید محتوا باشد.

اخبار امروزه نه تنها توسط خبرنگاران بلکه توسط شهروندان نیز تولید می‌شود. خبرهای مبتنی بر شهروندی در رسانه‌های دیجیتال گاه جریان اصلی خبر در رابطه با یک موضوع را شکل می‌دهد، به‌طوری‌که به‌طور فزاینده‌ای شاهد محتوای تولید شده توسط کاربر (که ممکن است شامل دیپ‌فیک باشد) هستیم و بازنشر آن در تلویزیون یا کانال‌های اجتماعی سازمان‌های بزرگ، که اگر صحت سنجی را کاهش دهیم، ممکن است پیامدهای جدی را دربر داشته باشد. به‌عنوان‌مثال، یک ویدئو یا عکس جعلی غیرواقعی (تولید شده با دیپ فیک) که توسط شهروندان ایجاد شده اگر در سازمان‌های خبری بزرگ به‌صورت رسمی منتشر شود با انگیزه‌های متفاوت از سیاسی و اجتماعی تا موثر بر اقتصاد و بازار با اخلاق حرفه‌ای سازمانی در تضاد است [118]. به‌این‌ترتیب، ظهور دیپ‌فیک‌ها تاثیر قابل‌توجهی بر رسانه‌ها می‌گذارد، زیرا مطمئنا از تولید آن بر اساس دستور کار سازندگان آن سوء استفاده می‌شود.

[1] Twitter (X)

[2] Face Book

۳-۳- دیدگاه بازنمایی رسانه‌ها

نمایش کلی هوش مصنوعی و قابلیت‌های آن در رسانه‌های دیجیتال به دلیل اینکه نشان می‌دهد که چگونه چنین پیشرفت‌های تکنولوژیکی و تاثیرات آن به مردم منتقل می‌شود موردتوجه همگان است [119]. هوش مصنوعی در بیشتر فرهنگ‌ها عمدتا به‌عنوان روبات‌ها یا سیستم‌های رایانه‌ای بسیار پیشرفته که دارای منطق انسان‌مانند و قابلیت‌های مکالمه هستند، شناخته می‌شود.

این امر را در داستان‌ها، تصاویر و فیلم‌ها مشاهده می‌کنیم که درنهایت فرهنگ روبات‌ها را در دهه‌های گذشته ایجاد کرد. بااین‌حال، با سیستم‌های هوش مصنوعی مدرن، تجسم فیزیکی آن دیگر به‌عنوان یک اصل مطرح نیست، چون هوش مصنوعی در فرآیندهای زندگی روزمره هم وجود دارد و حتی به‌طور مستقیم قابل تشخیص نیست، به‌عنوان‌مثال، تعامل با دستیارهای صوتی در تلفن‌های هوشمند. درک هوش مصنوعی در عموم مردم ب این است که روبات‌هایی (گاهی اوقات انسان‌نما) هستند که از کلیشه‌ها یا دستورات پیروی می‌کنند و اغلب ناجی و پیرو هستند و گاهی هم تصمیم‌گیرنده یا شرور [120].

چنین کلیشه‌هایی ساختار ذهن ما از جهان، ارزش‌ها و تجربیات ما می‌باشد و قرار دادن روبات‌ها در یک دسته‌بندی خاص را ارایه می‌کند. بنابراین، انتظارات و تعاملات ما با آن‌ها و اینکه امروزه رسانه‌ها چگونه هوش مصنوعی را نشان می‌دهند اغلب مغرضانه و نامتعادل است، و همین امر در مورد دیپ فیک نیز صدق می‌کند.

هوش مصنوعی عمومی اغلب در رسانه‌ها به‌عنوان ترکیبی از سرگرمی و ترس [121] به تصویر کشیده می‌شود. به‌عنوان‌مثال، دیپ‌فیک‌ها نمایانگر یک چیز که توسط یک ماشین (سیستم خودکار کامپیوتری) هستند، ممکن است یک اثر شگفت‌انگیز ایجاد کنند، تصور کنید چه اتفاقی می‌افتد زمانی که ماشین‌ها قادر به تقلید و کپی از هرکسی شوند از موضوع از بین رفتن مشاغل مختلف تا تحت تاثیر شدید قرار گرفتن آن‌ها، فراموش نکنیم اعتصاب بلندمدت نویسندگان آثار سینمایی و کتب در سال ۲۰۲۲ تا ۲۰۲۴ را که تاثیر زیادی بر اقتصاد سینما نیز داشت. این مسایل بر شناخت ما از هوش مصنوعی تاثیر بگذارند و به سوی دیدگاه خاصی درباره دیپ فیک‌ها و فناوری‌های مرتبط هدایت کنند.

به‌عنوان‌مثال، درحالی‌که دیدگاه علمی-تخیلی ممکن است نگرش‌های منفی به ربات‌های قاتل را در بخشی از جامعه‌ای که متقاضی بسیاری از داستان‌های علمی-تخیلی هستند را تشدید کند، این موضوع را نمی‌توان برای عموم جامعه ادعا کرد [122].

فرضیه دره غیرعادی[1] [123]، نشان می‌دهد که محصولات نهایی که شباهت زیادی به رفتار انسان دارند (اما نه دقیقا) می‌توانند موجب به وجود آمدن احساسات عجیب و غریبی شوند. بااین‌حال، نشانه‌ای وجود دارد که داستان‌های علمی تخیلی می‌تواند وهم‌آوری روبات‌ها را کاهش دهد [124]، چیزی که در آینده ممکن است شاهد آن برای محصولات دیپ‌فیک نیز باشیم. دیپ‌فیک‌ها را می‌توان با استفاده از هوش مصنوعی تولید کرد، با توجه به فرهنگ روباتی توسعه یافته طی دهه‌ها، می‌توان هوش مصنوعی و دیپ‌فیک‌ها را به ربات‌ها یا ماشین‌هایی مرتبط کرد که برای تولید محتوای جدید و فریب انسان‌ها به اندازه کافی هوشمند هستند. بااین‌حال، دیپ‌فیک‌ها صرفا به استفاده از الگوریتم‌های پیچیده برای محتوا متکی هستند تا ویدیو، صدا و تصاویر با کیفیت بالا و جعلی ایجاد کنند، اما به‌طور مستقل و هوشمندانه عمل نمی‌کنند. پیشرفت‌های سریع در فناوری در پشت دیپ‌فیک‌ها، چنین مسایلی نیاز به بازنگری و درک بهتر دارند.

۴-۳- دیدگاه مخاطبان رسانه

رسانه‌های دیجیتال مطمئنا نحوه تعامل افراد در همه سطوح چه تعامل با یکدیگر چه با رسانه‌ها و فضای مجازی را تغییر داده‌اند و اکنون با افزایش هوش مصنوعی، به نظر می‌رسد که چنین روابطی دوباره به‌طور قابل توجهی تحت تاثیر قرار خواهند گرفت. جامعه این اثر در نتیجه‌ی قرار گرفتن در معرض نفوذ رسانه‌های جمعی است.

---

[1] Uncanny valley

با در نظر گرفتن قابلیت‌های دیپ‌فیک که ایجاد محتوای جدید و انتشار آسان آن از طریق رسانه‌های دیجیتال و شبکه‌های اجتماعی است، بدیهی است که دیپ‌فیک یک تغییردهنده بازی خواهد بود و تاثیری آن چندوجهی و چندجانبه خواهد بود. به‌عنوان‌مثال، افراد هنگام تعامل با هوش مصنوعی رفتارهای متفاوتی دارند و همچنین مشاهده شده است که هوش مصنوعی به‌عنوان یک مشاور می‌تواند آسیب‌زننده هم باشد، همچنین می‌دانیم رسانه‌های اجتماعی بدون هوش مصنوعی هم پتانسیل تاثیر روی افراد را دارند که این امر باوجود فراگیری زیاد و ورود هوش مصنوعی، باعث دخالت آن در کنش‌های اجتماعی، مانند جنبش‌های اجتماعی هم می‌شود [125].

امروزه، بیش از هر زمان دیگری، یک فرد می‌تواند به‌صورت قابل‌توجه تاثیرگذار باشد، زیرا اگر پیام او در فضای مجازی منتشر شود، میلیون‌ها نفر آن را می‌بینند که بر آن‌ها تاثیر می‌گذارد. دیپ‌فیک‌ها، به‌ویژه با موضوعات بحث‌برانگیز، دارای چنین پتانسیلی برای اثرگذاری هستند، زیرا افراد ممکن است با توجه به حساسیت زمانی یا موضوعی موضوع، تمایل کمتری به بررسی صحت آن‌ها نداشته باشند. به‌عنوان‌مثال، پیام‌های سیاسی در زمان درگیری بین ملت‌ها می‌تواند به‌طور غیرقابل کنترلی منتشر شود و واکنش‌های غیرمنطقی و احساسی را به دنبال داشته باشد. طرفداری ممکن است جنبه دیگری باشد که تحت تاثیر قرار می‌گیرد. امروزه همگرایی رسانه‌ها، فناوری‌های جدید و بازاریابی رسانه‌ای همگی انواع جدیدی از طرفداران را در سمت مخاطبان ایجاد کرده‌اند دیپ‌فیک‌ها به‌طور بالقوه می‌توانند عاملی تاثیرگذار بر میزان طرفداری افراد از یک شخص یا جریان باشند. دیپ‌فیک‌ها این پتانسیل را دارند که جوامع را سریع تحت تاثیر قرار دهند و محتوای دیپ‌فیک تولید شده می‌تواند راحت‌تر طرفداران را جذب کند، چیزی که نشان می‌دهد رسانه دیگر ممکن است به‌تنهایی چنین فضاهایی را کنترل نکند.

۵-۳- دیدگاه جنسیتی

ظهور دیپ فیک را از جنبه جنسیتی نیز می‌توان بررسی کرد. فمینیسم را می‌توان یک جنبش رهایی‌بخش با هدف از بین بردن سلطه و ظلم در نظر گرفت. وقتی به‌طور اجتماعی به تاثیرات رسانه و موضوع فمینیسم نگاه می‌کنیم، وجود فلسفه‌ها، تفاسیر متفاوت از مفاهیم و منطق فمینیستی را در رسانه‌ها می‌بینیم. سوالی که مطرح می‌شود این است که تداخل دیپ فیک و فمینیسم چه می‌تواند باشد. ازآنجایی‌که فمینیسم معاصر، اغلب به‌عنوان فمینیسم هشتگ نیز شناخته می‌شود، به این صورت که به یکباره این هشتگ داغ[1] شده و مساله اصلی در شبکه‌های اجتماعی برای گفت‌وگو، اعلام نظر و گاهی اعتصابات می‌گردد، این مساله به صورت آنلاین و گاهی منحصرا از طریق پلتفرم‌های رسانه‌های اجتماعی اتفاق می‌افتد [125]، به شدت مستعد ساخت محتوای جعلی دیپ‌فیک است. نباید فراموش کرد که مطالب دیپ فیک اولیه که ظاهر شد عمدتا فیلم‌های جعلی مستهجن سلبریتی زنان را به تصویر می‌کشید. در این رابطه اگرچه اغلب در رسانه‌ها این بحث مطرح می‌شد که برای سرگرمی ایجاد شده است و عمدتا مخاطبان مرد را مورد خطاب قرار می‌دهد، اما آن‌ها در سایت‌های معروف پورنوگرافی در دسترس عموم مردم قرار می‌گیرند و به‌طور تاثیرگذاری به هویت و جایگاه اخلاقی افراد لطمه زده می‌شود. درحالی‌که چنین محتوایی غیرقانونی است و وب‌سایت‌های مختلف پس از شناسایی برای حذف آن اقدام می‌کنند که البته اغلب چنین اقداماتی خیلی دیر انجام می‌شوند یا کارآمد نیستند.

اکنون می‌توان ویدیوهای ظاهرا واقعی با صداهای منطبق ایجاد کرد و به راحتی بین مخاطبان در فضای مجازی منتشر کرد. به‌عنوان‌مثال، در حال حاضر از هر بیست‌وپنج آمریکایی یک نفر قربانی ویدئوهای غیراخلاقی انتقامی شده است [126]، چیزی که انتظار می‌رود با توجه به کیفیت بالا و همچنین سهولتی که در دسترسی به دیپ‌فیک‌ها وجود دارد، افزایش یابد.

در ادبیات موضوعی که بررسی شد عواملی مشاهده شده است که مزایای فمینیست‌ها را محدود می‌کند. به‌طور خاص مثلا فمینیست‌ها اشکال جدیدی از محرومیت از دسترسی به رسانه‌ها و تبلیغات را تجربه می‌کنند، زیرا شبکه‌های دیجیتال اجتماعی می‌توانند محلی برای ایجاد آسیب برای آن‌ها باشد. دیپ‌فیک‌ها پتانسیل افزایش چنین عدم اطمینان و محدودیت‌هایی را دارند، زیرا می‌توانند به راحتی محتوای تبعیض‌آمیز را ایجاد و منتشر کنند. لازم به ذکر است که جنبه‌های جنسیتی به معنای هدف قرار گرفتن زنان و مردان نیست البته می‌توان اقدامات مشابهی را علیه مردان نیز انجام داد. بااین‌حال، اکثریت موارد تاکنون علیه زنان و در نقش‌های خاص بوده است. به‌طورکلی، جنسیت باید به‌عنوان یک ساختار اجتماعی موردتوجه قرار گیرد، و جنبه‌های جنسیتی، در پیوند با دیپ فیک‌ها، باید در حوزه وسیع‌تر نظریه فمینیستی و مطالعات جنسیتی موردتوجه قرار گیرد.

[1] Trend

۶-۳- دیدگاه قانون و مقررات

برخی کشورها قوانین و چارچوب نظارتی دارند که مربوط به رسانه‌های دیجیتال و مسایل مربوط به آن‌ها است. انتشار اخبار جعلی، از جمله دیپ‌فیک، نیز از طریق همان فرآیندها موردتوجه و نظارت قرار می‌گیرد. به‌عنوان‌مثال، در ایالات‌متحده، قانون پیشنهادی منع مخرب دیپ‌فیک در سال ۲۰۱۸ یک جرم کیفری جدید مربوط به ایجاد یا توزیع محتوای جعلی که واقعی به نظر می‌رسد را درنظر می‌گیرد. علاوه بر این، یک قانون پاسخگویی در سال ۲۰۱۹ به منظور مبارزه با انتشار اطلاعات نادرست که از طریق فناوری‌های جدید و تغییرات یا ساخت ویدئوهای جعلی با دیپ‌فیک ساخته می‌شود تصویب شد [127].

بااین‌حال، حتی با وجود چنین اقدامات قانونی هوش مصنوعی و پیامدهای آن به‌خوبی درک نشده است و قوانین فعلی نمی‌توانند پیچیدگی هوش مصنوعی را مدیریت کنند یا بازدارندگی لازم را برای کاهش جرایم ایجاد کند، اگرچه بر این موضوع واقف هستیم که اجرای قانون یا رعایت آن بحثی فرهنگی در اجتماع است.

برخی اقدامات خطرات مرتبط با دیپ فیک را کاهش نمی‌دهد و باید گام‌های موثرتری برداشته شود، اقدامات غیر واقعی از نظر تاثیرگذاری انجام شده برای مثال، نیاز به قراردادن واترمارک[1] و برچسب‌گذاری واضح بر روی محتوای ساخته شده توسط دیپ‌فیک چیزی است که مطمئنا سازندگان دیپ‌فیک، به‌ویژه آن‌هایی که مقاصد غیراخلاقی دارند، از آن تبعیت نمی‌کنند. به‌این‌ترتیب، اثربخشی آن محدود است؛ پس بدون قانون‌گذاری مناسب از دیپ‌فیک‌ها، هم قوه مجریه که قانون را اجرا می‌کند و هم قوه قضاییه که قوانین را تفسیر می‌کند، با چالش‌هایی مواجه خواهند شد. همان‌طور که مشاهده می‌شود، زمینه کامل دیپ‌فیک‌ها به‌خوبی درک نشده است و این خطر وجود دارد که اولا پرداختن به آن به شیوه‌ای موثر اتفاق نیافتد، بلکه فقط به صورت سطحی باشد، ثانیا معرفی اقداماتی که پیامدهای آن‌ها به‌خوبی درک نشده است و تضعیف حقوق مدنی است که ممکن است به اثرات اجتماعی درازمدت منجر شود.

۷-۳- دیدگاه سیاسی

فناوری بر نحوه تفکر ما در مورد امور اجتماعی و سیاسی تاثیر می‌گذارد [128]. به خصوص با آخرین پیشرفت‌های هوش مصنوعی، نه تنها اخبار، عکس‌ها و ویدیوهای جعلی (*Deepfakes*) می‌توان ایجاد کرد، بلکه می‌توان بررسی‌های جعلی، متن‌های واقعی متقاعدکننده و حتی مکالمات را در به صورت بلادرنگ و در لحظه ایجاد یا ویرایش کرد [110].

تاثیرگذاری و متقاعد کردن گروه‌های سیاسی در حال حاضر از طریق رسانه‌های اجتماعی مانند توییتر و فیس‌بوک امکان‌پذیر است و پیام‌رسان‌هایی مانند واتس‌اپ، تلگرام و... مخاطبان جهانی را به صورت شبانه‌روزی و در لحظه باهم در ارتباط قرار می‌دهند. قدرت این رسانه‌ها قبلا در نقش آن‌ها در جنبش‌های اجتماعی اخیر مانند بهار عربی در برخی از کشورهای درحال‌توسعه نشان داده شده است [129]. اطلاعات نادرست در چنین رسانه‌هایی در برهه‌های زمانی حساس اهمیت بسیاری دارد و این ویدئوهای سیاسی جعلی ایجاد شده [130]، پتانسیل افزایش بی‌اعتمادی به اخبار در رسانه‌های اجتماعی را دارد. هم رسانه‌های اجتماعی و هم پیام‌رسان‌ها به‌عنوان "زمین حاصلخیز برای انتشار دیپ‌فیک‌ها با پیامدهای انفجاری برای سیاست" در نظر گرفته می‌شوند.

موضوع اخبار جعلی مساله‌ای مهم چندجانبه است، زیرا در برخی از کشورهای درحال‌توسعه، نفوذ رسانه‌های اجتماعی به قدری زیاد است که اغلب به‌عنوان منبع اصلی اطلاعات و همچنین بررسی حقایق در نظر گرفته می‌شود. مشکل این است که انتشار اخبار جعلی اعتماد به منابع قانونی مانند شبکه‌های رسمی اطلاع‌رسانی آن کشور را کاهش می‌دهد که می‌تواند اثرات مخربی بر وضعیت آرامش اجتماعی، سطح اعتماد مردم به دولت‌ها و پذیرش و اطاعت‌پذیری آن‌ها داشته باشد.

[1] Watermark, Logo

افراد پس از قرار گرفتن در معرض اخبارهای جعلی ممکن است از نظر احساسی واکنش نشان دهند یا به اقداماتی هدایت شوند که نظم عمومی کشور را تحت تاثیر قرار دهد. چنین اقداماتی منجر به تضعیف کارکردهای حیاتی می‌شود و مغایر با حقوق اساسی در دولت‌ها است [132]، دیپ فیک چنین تهدیدی را ایجاد می‌کند، زیرا پتانسیل این را دارد که نقش اساسی در کاهش اعتماد به نهادهای عمومی داشته باشد.

۴- بحث

عصر دیپ‌فیک که در آن محتوای هیجان‌انگیز، غیرصادقانه یا حتی ساختگی بیشتر از طریق رسانه‌های اجتماعی منتشر می‌شود، اکنون فرا رسیده است و رویکردهای سنتی تفکر که از طریق گفته‌های رایجی مانند دیدن یعنی باور کردن، من به آنچه می‌بینم اعتماد دارم و یک تصویر هزار کلمه ارزش دارد و مطالبی از این دست به چالش کشیده می‌شوند.

ساخت عکس و فیلم همیشه چالش‌برانگیز بوده و تنها با تلاش و تخصص قابل توجهی می‌توان به آن دست یافت، اما دیپ‌فیک‌ها دسترسی آسان به ایجاد ویدیوهای جعلی واقع‌گرایانه را نشان داده‌اند و به‌این‌ترتیب، جنبه‌های تولید رسانه به‌طور قابل‌توجهی هم بر «چگونگی» و هم بر «چیزی» تولید می‌شود متمرکز شده است. برای مواجهه درست با پدیده دیپ فیک، باید آن را در حوزه عمومی قرار دهیم، یعنی جایی که بیشتر اتفاق می‌افتد را بررسی کنیم تا بتوان اثرات آن را مشاهده کرد. ما به‌عنوان حوزه عمومی، محدوده زندگی اجتماعی را در نظر می‌گیریم که در آن می‌توان به‌صورت افکار عمومی آن را درنظر گرفت [132] که مبنای هنجاری برای یک دموکراسی نیز می‌باشد. در این زمینه، رسانه‌ها به‌عنوان بستری برای بحث‌های پرطرفدار در نظر گرفته می‌شوند و با دموکراسی و جامعه مرتبط هستند، زیرا افراد و گروه‌ها از طریق آن حمایت یا عدم حمایت خود از دیدگاهی خاص را بیان می‌کنند.

نقش رسانه‌ها در حکومت‌داری و دموکراسی‌سازی حیاتی است، زیرا جنبه‌هایی مانند جامعه و رسانه‌های اجتماعی (که در آن‌ها ممکن است از دیپ‌فیک استفاده شود) بر حوزه‌های کلیدی مانند فقر، نابرابری و به‌طورکلی جامعه تاثیر می‌گذارد. پلتفرم‌های رسانه‌های اجتماعی جایی هستند که دیپ‌فیک‌ها عمدتا توزیع می‌شوند و به‌این‌ترتیب، آن‌ها به بخشی جدایی‌ناپذیر از پو یایی پیچیده آن‌ها هم تبدیل می‌شوند.

به‌این‌ترتیب، دیپ‌فیک‌ها از زمانی که پلتفرم‌های اجتماعی اجتماعی بودن را مهندسی می‌کنند، پیامدهای بزرگی دارند [82]. مسایل جدید پیرامون شبکه‌های اجتماعی و گوشی‌های هوشمند کلیدی در عصر دیجیتال وجود دارد [133]. در این عصر، جعل در رسانه‌ها چیز جدیدی نیست، اما این واقعیت که هر فردی با مهارت‌های فنی پایین می‌تواند آن را به راحتی انجام دهد، چالش‌های جدیدی را ایجاد می‌کند و بر تعامل رسانه‌ها و جامعه تاثیر می‌گذارد. دیپ‌فیک‌ها یک تجلی قدرتمند فناوری جدید از پدیده شناخته شده‌ی ایجاد محتوای جعلی را تشکیل می‌دهد.

برای ارزیابی واقعی بودن عکس‌ها می‌توان از متخصصان فن‌آوری در این حوزه‌ها استفاده کرد [134]. ابزارهای جدید مبتنی بر فناوری توسعه داده شوند که به شناسایی دیپ‌فیک‌ها کمک کند.

علاوه بر این، درحالی‌که ممکن است بخشی از کشورهای جهان با فناوری‌های پیشرفته‌تر آشنا باشد و رسانه‌های مستقل‌تری داشته باشد، بسیاری از کشورهای درحال‌توسعه، به دلیل شکاف دیجیتالی یا فقدان تکثر رسانه‌های مستقل، شهروندانی دارند که اغلب در حال تقلب هستند. یک نظرسنجی [135] مشخص کرد که بخش بزرگی از علاقه‌مند شدن به اخبار جعلی به دلیل تمایل عمومی به پذیرش بیش از حد ادعاهای ضعیف به‌عنوان یک مرجع عمومی است.

عدم آ گاهی، به‌عنوان یک مشکل در شناسایی اخبار جعلی به‌طورکلی دیده می‌شود [136]. بررسی حقایق در شناسایی اخبار جعلی بسیار مهم است، و افرادی که منابع خود را بررسی می‌کنند، شانس کمتری برای گول خوردن از مطالب جعلی و انتشار بیشتر آن دارند. وب‌سایت‌هایی هم وجود دارند که حقایق را بررسی می‌کنند، و افراد با شک و تردید می‌توانند اطلاعات دریافت‌شده را قبل از اعتماد به آن یا انتشار آن در رسانه‌های اجتماعی بررسی کنند. بااین‌حال، با دیپ فیک، همه‌چیز چالش‌برانگیزتر و پیچیده تر است، به خصوص به دلیل ماهیت واقع‌گرایانه‌ی آن‌ها. اگر افراد از قبل نسبت به متون، عکس‌ها و سایر مطالب نگاهی انتقادی داشته باشند، سواد رسانه‌های اجتماعی و سواد کلی نسبت به اخبار جعلی، ممکن است به کاربران این امکان را بدهد که گرفتار دیپ‌فیک نشوند. وجود دیپ‌فیک‌ها همچنان چالش‌برانگیز است و می‌تواند موردتوجه قرار گیرد.

در عصر حاضر سواد خبری و سواد فناوری شهروندان به‌عنوان یک مهارت اساسی در نظر گرفته می‌شود و استدلال استنتاج درست در محیط‌های رسانه‌های اجتماعی به‌عنوان یک مهارت انتقادی در نظر گرفته می‌شود که به‌عنوان‌مثال به دانش آموزان باید آموزش داده شود [137]. سواد دیجیتال می‌تواند به اتخاذ رویکرد سالم‌تری در رسانه‌های اجتماعی کمک کند و به‌عنوان عاملی برای مبارزه با انتشار اطلاعات جعلی، از جمله دیپ‌فیک‌ها، عمل کند؛ بنابراین، تعامل موثر با سازمان آموزش‌و‌پرورش اهمیت دارد تا سواد رسانه‌های دیجیتال از دوران کودکی افزایش یابد. بااین‌حال، حتی افراد باسواد در زمینه فناوری و رسانه‌های اجتماعی، وقتی با موارد دیپ فیک مواجه می‌شوند، ممکن است تحت تاثیر قرار بگیرند. آموزش و ارتقای مهارت شهروندان تنها قدم موثر نیست، علاوه بر این، خبرگزاری‌ها و رسانه‌ها باید استانداردهای باکیفیتی را رعایت کنند و این امر ممکن است از طریق اقدامات قانونی برای مجازات سنگین برای انتشار اطلاعات نادرست رسانه‌ها نیز اجرا شود.

شفافیت اخبار آنلاین رسانه‌ها و بررسی منابع باید امری روتین و عادی باشد. علاوه بر این، ابزارهای جدیدی مورد نیاز است که هم شهروندان و هم خبرنگاران را قادر می‌سازد صحت اطلاعات را شناسایی و بررسی کنند و منبع آن را مشخص کنند.

باید ترکیبی از تجزیه‌وتحلیل انسان‌محور و همچنین هوشمند برای سنجش صحت اطلاعات وجود داشته باشد، به‌عنوان‌مثال، از طریق رویکردهای هوش مصنوعی [96] که بتواند به‌صورت لحظه‌ای و بلادرنگ عمل کند باید برای شناسایی دیپ‌فیک‌ها استفاده شود.

دیپ‌فیک‌ها پیامدهای سنگین و تاثیرگذاری دارند به‌ویژه در دوره‌ای که اطلاعات به راحتی در شبکه‌های اجتماعی، در سراسر جهان ارتباط منتشر می‌شوند. شبکه‌های اجتماعی زمینه مناسبی برای انتشار دیپ‌فیک‌ها هستند، به‌عنوان‌مثال، حوزه سیاست این پتانسیل را دارد که با تاثیر ویژه مثلا در فرآیند انتخابات فرآیندهای اجتماعی را مختل کند. مثلا زمانی که فرد یک عکس جعلی را با پیامی که به‌صورت جعلی ساخته شده، به منظور ایجاد یک نکته منفی یا مثبت در جامعه باهم منتشر کند. دیپ فیک می‌تواند شکل جدیدی از جنگ روانی معاصر و ابزار دستکاری افکار فردی یا گروهی در جامعه باشد.

در روزنامه‌نگاری خبری سیاسی، یعنی پوشش رسانه‌های خبری در حوزه سیاست، روزنامه‌نگاران سعی می‌کنند تا کنترل داستان‌های سیاسی را در دست داشته باشند تا اینکه به‌طور منفعلانه درباره آنچه توسط افراد سیاسی تبلیغ می‌شود گزارش دهند [17]. برای انجام این کار، آن‌ها باید قادر باشند تا در دام محتوای دیپ‌فیک تولید شده توسط سهامدارانی که به‌طور مستقیم یا غیرمستقیم از چنین افراد سیاسی حمایت می‌کنند، نیفتند. به‌عنوان‌مثال، در زمینه جنبش‌ها و فعالیت‌های اجتماعی ترکیب یک ویدیوی دیپ فیک با یک پیام جدی که متناسب با یک برنامه سیاسی خاص باشد، بر مردم تاثیر خواهد گذاشت. در گرماگرم اقدامات جنبش، از چنین ویدئوهایی می‌توان برای بی‌اعتبار ساختن مخالفان، ایجاد عکس‌های نزدیک به واقعیت که خشم شهروندان را برانگیزد تا زمان فاش شدن آن، واقعیتی موقت ولی جعلی ایجاد می‌کند، استفاده کنند تا به اهداف خود برسند. این نکته را درنظر داشته باشید که اصولا یک شایعه سریع‌تر و بیشتر از تکذیبیه آن منتشر و دیده خواهد شد.

موضوع اعتماد به رسانه‌ها، یک چالش بزرگ است، حذف اعتماد از اخبار، تصاویر، ویدئوها اساسا اعتماد به تعاملات و ساختارهای اجتماعی را از بین می‌برد و عملا درب را برای شک در همه‌جا، حتی در موارد مشروع، باز می‌گذارد. بنابراین، یا اطلاعات نادرست، می‌تواند بر جامعه و فرآیندهای آن تاثیر بگذارد [130]. زوال حقیقت مصادف است با زوال اعتماد [16] که نگرانی‌های جدیدی را ایجاد می‌کند، زیرا جامعه دیگر نمی‌تواند درک درستی از واقعیت را به اشتراک بگذارد و بر اساس آن عمل کند.

بررسی این موضوع که چگونه فناوری بخشی از فرآیند است و چگونه فرآیندهای جاری در حوزه دیجیتال را تحت تاثیر قرار می‌دهد و تحت تاثیر قرار می‌گیرد، ضروری است. رسانه‌ها و دیپ فیک‌ها درحالی‌که دیپ‌فیک‌ها در این مرحله اولیه مستقیما انسان‌ها را هدف قرار می‌دهند، در آینده ممکن است این‌طور نباشد.

اتکای زیاد به سیستم‌های فیزیکی تکنولوژی مجهز به هوش مصنوعی، به‌عنوان‌مثال، خودروهای خودران، ممکن است منجر به ایجاد دیپ‌فیک‌هایی شود که ماشین‌ها و به‌طور غیرمستقیم انسان‌ها را هدف قرار می‌دهند. باز کردن قفل خودروی خودران با صدای دیپ‌فیک، تغییر رفتار آن با نمایش تصاویر دیپ‌فیک به سنسورهای آن و موارد دیگر می‌تواند بر رفتار طراحی شده خودرو تاثیر بگذارد و آن را به سمت تصمیم‌های غیرقابل پیش‌بینی

مثل ترمز ناگهانی یا مواردی از این دست سوق دهد که می‌تواند به انسان آسیب برساند. تحقیقات تجربی بیشتری در این زمینه‌ها مورد نیاز است، که همچنین به چارچوب‌های نظری مناسب مرتبط باشد و از آن‌ها پشتیبانی کند.

تهدیدات ناشی از دیپ فیک باید از طریق ترکیبی از فناوری، مقررات و آموزش موردتوجه قرار گیرد. آگاهی شهروندان همیشه برای پیشبرد تغییرات سیاسی و اجتماعی مبتنی بر اصلاحات موردتوجه است، اساسا آگاهی جامعه یکی از مواردی است که حکم‌رانان برای پو یایی جامعه و بالارفتن سواد رسانه‌ای و اجتماعی باید موردتوجه قرار دهند. تحقیقات اخیر [12]، [32] راه‌حل‌های مبتنی بر هوش مصنوعی را به تصو یر می‌کشد که می‌توانند دیپ‌فیک‌ها را شناسایی کنند، از جمله در بسیاری از موارد نرم‌افزاری که برای ایجاد آن‌ها استفاده شده است.

دیگران فناوری‌های تکمیلی مانند بلاک چین[1] [87] را به منظور پیوند دادن و یدئوها به نهادهای قابل اعتماد یا معتبر پیشنهاد کرده‌اند. درحالی‌که راه‌حل‌های فنی برای شناسایی، تایید، و حذف این‌گونه جعلی‌ها در حال انجام است [71]، [87]، [96] مشکل صرفا یک مشکل تکنولوژیکی نیست، بلکه اعتماد به فرآیندها و ذینفعان است، به‌عنوان‌مثال، به روزنامه‌نگاری که مسئولانه عمل می‌کند. دیپ فیک و فناوری زیربنایی آن، نه تنها تهدید، بلکه فرصت‌هایی را نیز به همراه دارد. الگوریتم‌های هوش مصنوعی مورد استفاده در دیپ فیک دارای طیف گسترده‌ای از قابلیت‌ها هستند و می‌توانند متن، صدا، و یدئو، آثار هنری و... جدید ایجاد کنند. علاوه بر این، حتی فناوری هسته دیپ فیک نیز می‌تواند اثرات مفیدی در شبیه‌سازی و آموزش پرسنل به صورت سفارشی داشته باشد. همچنین کاربردهایی، در ارایه قانونی محتوا و تاکتیک‌های فریب علیه جنایتکاران، تروریست‌ها و سایر دشمنانی که علیه منافع عمومی عمل می‌کنند.

۵- نتیجه‌گیری

این بررسی سیستماتیک ادبیاتی (*SLR*) روش‌های مختلف برای شناسایی دیپ فیک که در ۱۱۲ مطالعه را ارایه می‌دهد. ما تکنیک‌های پایه‌ای را ارایه داده و در این کار کارایی مدل‌های شناسایی مختلف را بررسی کردیم. خلاصه کلی این مطالعه به شرح زیر است:

۱. روش‌های مبتنی بر یادگیری عمیق به‌طور گسترده در شناسایی دیپ فیک استفاده می‌شوند و موثرترین روش‌ها این الگوریتم‌ها هستند.
۲. بیشترین تحلیل و آزمایش برای تشخیص دیپ‌فیک بر روی صورت و چشم اتفاق افتاده است.
۳. مدل‌های یادگیری عمیق (عمدتا *CNN*) درصد قابل توجهی از تمام مدل‌ها را تشکیل می‌دهند.
۴. پرکاربردترین معیار برای شناسایی دیپ‌فیک دقت یا *Accuracy* است.

نتایج آزمایشات نشان می‌دهد که تکنیک‌های یادگیری عمیق در شناسایی دیپ فیک موثر هستند. علاوه بر این، می‌توان گفت که به‌طورکلی، مدل‌های یادگیری عمیق بهتر از مدل‌های غیر عمیق عمل می‌کنند. با پیشرفت سریع در فناوری چندرسانه‌ای زیرساختی و گسترش ابزارها و برنامه‌ها، تشخیص دیپ فیک هنوز با چالش‌های زیادی روبرو است. امیدواریم که این *SLR* منبع ارزشمندی برای جامعه تحقیقاتی در توسعه روش‌های شناسایی و راهکارهای پاسخگو فراهم کند.

درک صحیح دیپ‌فیک در رسانه‌های دیجیتال مدرن و همچنین فرآیندهایی که بر آن تاثیر می‌گذارد و پیامدهای کلی آن، چالش‌برانگیز تلقی می‌شود؛ بنابراین، باید از زوایای متعددی مورد بررسی قرار گیرند. بااین‌حال، برای انجام این کار، باید ابعاد مناسبی تعریف شود (که امروزه عمدتا چنین نیست) و این‌ها باید برای ثبت همه عوامل دخیل در این مساله کافی باشد. این کار فقط برخی از جنبه‌های سطح بالا را آشکار کرده است و تحقیقات بسیار عمیق‌تری لازم است تا سطوح دیگر این تکنولوژی به همراه پیش‌بینی‌هایی از وضعیت آینده شفاف گردد. این خطر ذاتی وجود دارد که جامعه دیگر نتواند به‌موقع جنبه‌های واقعی و جعلی را به‌طور معتبر تشخیص دهد، که ممکن است اعتماد به ذینفعان، فرآیندها و روزنامه‌نگارها، رسانه‌ها و شبکه‌های رسمی و اجتماعی را کاهش دهد و شعار «همه‌چیز جعلی است» ممکن است غالب شود. درحالی‌که راه‌حل‌های فنی برای شناسایی، تایید، و حذف چنین جعل‌هایی مورد نیاز است، مشکل کاملا فنی نیست، بلکه باید شامل اقدامات نظارتی قانونی و همچنین جنبه‌های آموزشی برای افزایش سواد تکنولوژی و عمومی کاربران باشد.

[1] Blockchain

۱-۵- تحقیقات آتی

این مقاله ضمن بررسی و مقایسه مطالعات در حوزه تشخیص دیپ‌فیک، به دنبال آشکارسازی و هشدار جدی پیرامون تاثیر محتوای جعلی در رسانه‌ها است، بررسی این موارد در سطح جهان و تطبیق آن با کشور ایران می‌تواند از مسایلی باشد که باید به‌طور جدی توسط پژوهشگران و متخصصان پیگیری شود.

دیپ‌فیک یک فناوری قدرتمند را در عصر نوظهور هوش مصنوعی نشان می‌دهد و همان‌طور که در مورد تمام فناوری‌های تغییر پارادایم وجود دارد، استفاده از آن تنها با قابلیت‌های آن تعیین نمی‌شود، بلکه چارچوب نظارتی، اخلاق، فرهنگ و سایر هنجارهای اجتماعی نیز نقش تعیین‌کننده‌ای دارند؛ بنابراین، به‌عنوان راه‌های تحقیقات آتی، می‌توان جنبه‌های متعددی را در نظر گرفت که در این کار به آن‌ها اشاره شده است، اما باید به‌طور عمیق‌تری به آن‌ها پرداخت. چنین جنبه‌هایی شامل رویکردی سخت‌گیرانه به رابطه بین دیپ فیک و جامعه و همچنین تاثیرات آن است. این باید شامل نحوه نمایش آن‌ها و همچنین رفتار آن‌ها در طول زمان باشد. علاوه بر این، یک رویکرد متقاطع دقیق که هویت‌ها را با جزییات پوشش می‌دهد، به‌عنوان‌مثال، جنسیت، نژاد، طبقه، تمایلات جنسی، ناتوانی، و نقش و تاثیر آن‌ها برای تبعیض و بی‌عدالتی اجتماعی مفید خواهد بود.

تعامل با رسانه، فرهنگ و جامعه چالش‌برانگیز است و در این زمینه، پژوهش تجربی، همراه با موقعیت خوب در چارچوب‌های نظری، وجود ندارد. تحقیقات تجربی که چارچوب‌های نظری ملموس را با استفاده از دیپ‌فیک‌ها و تاثیر در موارد استفاده پیوند می‌دهد و زمینه‌های نظری پیشنهادی را تایید یا رد می‌کند، ضروری است؛ همچنین باید تلاش‌ها برای شناسایی موارد جعلی مبتنی بر فناوری، به‌عنوان‌مثال، ویدئو، صدا، متن و اینکه چگونه این تلاش‌ها ممکن است به ابزارهایی منجر شود که می‌تواند توسط ذینفعان، مانند روزنامه‌نگاران، شهروندان و ... استفاده شود، انجام شود. رویکردهایی که اعتماد به دیجیتال را افزایش می‌دهد. منابع رسانه‌ای و فرآیندهای وابسته نیز مورد نیاز است و تحقیقات می‌تواند به ساخت پلتفرم‌ها، خدمات و ابزارهای دنیای واقعی که آن را قادر می‌سازد اختصاص داده شود.

ازآنجایی‌که دیپ فیک با چندین حوزه از زندگی مدرن تلاقی می‌کند، بررسی جنبه اخلاقی آن و نیز حوزه‌های مرتبط با ایمنی و امنیت در بافت جامعه مهم است. درنهایت، تحقیقات باید به جنبه‌های آموزشی منتهی شود، به‌عنوان‌مثال، شهروندان و همچنین افرادی که در حکومت‌داری، مانند قوه مقننه، مجریه و قضاییه دخیل هستند، اختصاص یابد. درک تقاطع رسانه‌های دیجیتال و هوش مصنوعی، در مورد دیپ فیک‌ها، و تاثیر آن بر جامعه مدرن ضروری است تا این فناوری به درستی در دسترس قرار گیرد و چالش‌هایی که ایجاد می‌کند به‌طور موثر موردتوجه قرار گیرد. از بحث‌ها مشخص می‌شود که تلاقی رسانه‌های دیجیتال و دیپ‌فیک، تاثیرات متعددی بر افراد و جامعه دارد که پرداختن به آن‌ها حایز اهمیت است.

مشارکت نویسندگان

سهیل فاخری ایده‌پردازی و طراحی کلی پژوهش را بر عهده داشت و مسئولیت نگارش بخش‌های اصلی مقاله را بر عهده گرفت. اعظم‌السادات نوربخش در جمع‌آوری منابع و مرور ادبیات پژوهش همکاری داشته و در تحلیل داده‌ها نقش داشت. محمدرضا یمقانی در انجام تحلیل‌های فنی و بررسی روش‌های تشخیص دیپ‌فیک و همچنین در ویرایش نهایی مقاله مشارکت نمود. همه نویسندگان نسخه نهایی مقاله را مطالعه و تأیید کردند.

منابع مالی

نویسندگان از هیچ منبع مالی جهت تدوین پژوهش استفاده ننموده اند.

تعارض با منافع

نویسندگان این پژوهش می‌دارند که هیچ تضادی در منافع در مورد انتشار این نسخه وجود ندارد، همه نویسندگان، نسخه نهایی ارسال شده را مشاهده و تایید کرده‌اند. همچنین، نویسندگان تضمین میکنند که این پژوهش، اثر اصلی آنها بوده، قبال چاپ نشده و در حال حاضر تحت انتشار نمی‌باشد.

منابع

[1] Chesney, R., & Citron, D. (2019). Deepfakes and the new disinformation war: the coming age of post-truth geopolitics. *Foreign aff.*, *98*, 147. https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/fora98&section=18

[2] Westerlund, M. (2019). The emergence of deepfake technology: a review. *Technology innovation management review*, *9*(11). https://timreview.ca/article/1282

[3] Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: trick or treat? *Business horizons*, *63*(2), 135–146. https://doi.org/10.1016/j.bushor.2019.11.006

[4] Fletcher, J. (2018). Deepfakes, artificial intelligence, and some kind of dystopia. *Theatre journal*, *70*(4), 455–471. https://doi.org/10.1353/tj.2018.0097

[5] Peele, J. (2018). *You won't believe what obama says in this video*. https://www.buzzfeed.com/watch/video/52602

[6] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … & Bengio, Y. (2014). *Generative adversarial nets* [Presentation]. Advances in neural information processing systems (p. 27). https://proceedings.neurips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

[7] YouTube. (2019). *Taxi driver starring Al Pacino [DeepFake]*. https://www.youtube.com/watch?v=9NkKj0aNB0s

[8] GitHub. (2020). *Avatarify: avatars for zoom, Skype and other video-conferencing apps.* https://github.com/alievk/avatarify

[9] Topsakal, O., Dobratz, E. J., Akbas, M. I., Dougherty, W. M., Akinci, T. C., & Celikoyar, M. M. (2023). Utilization of machine learning for the objective assessment of rhinoplasty outcomes. *IEEE Access*, *11*, 42135-42145.

[10] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401-4410). IEEE.

[11] Patrini, G., Cavalli, F., & Ajder, H. (2018). *The state of deepfakes: reality under attack.* https://regmedia.co.uk/2019/10/08/deepfake_report.pdf

[12] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2018). *Faceforensics: a large-scale video dataset for forgery detection in human faces*. https://doi.org/10.48550/arXiv.1803.09179

[13] Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., … & Theobalt, C. (2018). Deep video portraits. *ACM transactions on graphics (tog)*, *37*(4), 1–14. https://doi.org/10.1145/3197517.3201283

[14] Chan, C., Ginosar, S., Zhou, T., & Efros, A. A. (2019). Everybody dance now. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 5933-5942). IEEE.

[15] Smidi, A., & Shahin, S. (2017). Social media and social mobilisation in the middle east: a survey of research on the arab spring. *India quarterly*, *73*(2), 196–209. https://doi.org/10.1177/0974928417700798

[16] Chesney, B., & Citron, D. (2019). Deep fakes: a looming challenge for privacy, democracy, and national security. *California law review*, *107*, 1753. https://doi.org/10.15779/Z38RV0D15J

[17] Matern, F., Riess, C., & Stamminger, M. (2019). Exploiting visual artifacts to expose deepfakes and face manipulations. *2019 IEEE winter applications of computer vision workshops (WACVW)* (pp. 83–92). IEEE.

[18] Ciftci, U. A., Demir, I., & Yin, L. (2020). Fakecatcher: detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence* (pp. 1-17). IEEE.

[19] Bonomi, M., Pasquini, C., & Boato, G. (2021). Dynamic texture analysis for detecting fake faces in video sequences. *Journal of visual communication and image representation*, *79*, 103239. https://doi.org/10.1016/j.jvcir.2021.103239

[20] Guarnera, L., Giudice, O., & Battiato, S. (2020). Fighting deepfake by exposing the convolutional traces on images. *IEEE access*, *8*, 165085–165098. https://ieeexplore.ieee.org/abstract/document/9189772/

[21] Li, X., Lang, Y., Chen, Y., Mao, X., He, Y., Wang, S., … & Lu, Q. (2020). Sharp multiple instance learning for deepfake video detection. *Proceedings of the 28th ACM international conference on multimedia* (pp. 1864–1872). Association for Computing Machinery.

[22] Yang, X., Li, Y., & Lyu, S. (2019). Exposing deep fakes using inconsistent head poses. *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 8261–8265). IEEE.

[23] Sahla Habeeba, M. A., Lijiya, A., & Chacko, A. M. (2021). Detection of deepfakes using visual artifacts and neural network classifier. *Innovations in electrical and electronic engineering: proceedings of ICEEE 2020* (pp. 411–422). Springer.

[24] Zhang, X., Karaman, S., & Chang, S.-F. (2019). Detecting and simulating artifacts in gan fake images. *2019 IEEE international workshop on information forensics and security (WIFS)* (pp. 1–6). IEEE.

[25] Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2018). Learning rich features for image manipulation detection. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1053–1061). IEEE.

[26] Chugh, K., Gupta, P., Dhall, A., & Subramanian, R. (2020). Not made for each other-audio-visual dissonance-based deepfake detection and localization. *Proceedings of the 28th ACM international conference on multimedia* (pp. 439–447). Association for Computing Machinery.

[27] Hernandez-Ortega, J., Tolosana, R., Fierrez, J., & Morales, A. (2020). *Deepfakeson-phys: deepfakes detection based on heart rate estimation*. https://doi.org/10.48550/arXiv.2010.00400

[28] Fernandes, S., Raj, S., Ortiz, E., Vintila, I., Salter, M., Urosevic, G., & Jha, S. (2019). Predicting heart rate variations of deepfake videos using neural ode. *Proceedings of the ieee/cvf international conference on computer vision workshops* (p. 0). IEEE.

[29] Qi, H., Guo, Q., Juefei-Xu, F., Xie, X., Ma, L., Feng, W., … & Zhao, J. (2020). Deeprhythm: exposing deepfakes with attentional visual heartbeat rhythms. *Proceedings of the 28th ACM international conference on multimedia* (pp. 4318–4327). Association for Computing Machinery.

[30] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). Mesonet: a compact facial video forgery detection network. *2018 IEEE international workshop on information forensics and security (WIFS)* (pp. 1–7). IEEE.

[31] Kawa, P., & Syga, P. (2020). *A note on deepfake detection with low-resources*. https://doi.org/10.48550/arXiv.2006.05183

[32] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: learning to detect manipulated facial images. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1–11). IEEE.

[33] Khodabakhsh, A., & Busch, C. (2020). A generalizable deepfake detector based on neural conditional distribution modelling. *2020 international conference of the biometrics special interest group (biosig)* (pp. 1–5). IEEE.

[34] Chollet, F. (2017). Xception: deep learning with depthwise separable convolutions. *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1251–1258). IEEE.

[35] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708). IEEE.

[36] Li, Y., Chang, M.-C., & Lyu, S. (2018). In ictu oculi: exposing ai created fake videos by detecting eye blinking. *2018 IEEE international workshop on information forensics and security (WIFS)* (pp. 1–7). IEEE.

[37] Li, Y., & Lyu, S. (2018). *Exposing deepfake videos by detecting face warping artifacts*. https://doi.org/10.48550/arXiv.1811.00656

[38] Haliassos, A., Vougioukas, K., Petridis, S., & Pantic, M. (2021). Lips don't lie: a generalisable and robust approach to face forgery detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5039–5049). IEEE.

[39] Kukanov, I., Karttunen, J., Sillanpää, H., & Hautamäki, V. (2020). Cost sensitive optimization of deepfake detector. *2020 asia-pacific signal and information processing association annual summit and conference (APSIPA ASC)* (pp. 1300–1303). IEEE.

[40] Singh, A., Saimbhi, A. S., Singh, N., & Mittal, M. (2020). DeepFake video detection: a time-distributed approach. *SN computer science, 1*(4), 212. https://doi.org/10.1007/s42979-020-00225-9

[41] Ganiyusufoglu, I., Ngô, L. M., Savov, N., Karaoglu, S., & Gevers, T. (2020). *Spatio-temporal features for generalized detection of deepfake videos*. https://doi.org/10.48550/arXiv.2010.11844

[42] Wang, X., Yao, T., Ding, S., & Ma, L. (2020). Face manipulation detection via auxiliary supervision. *Neural information processing: 27th international conference, ICONIP 2020, part I 27* (pp. 313-324). Springer International Publishing.

[43] Zhu, X., Wang, H., Fei, H., Lei, Z., & Li, S. Z. (2021). Face forgery detection by 3d decomposition. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2929–2939). IEEE.

[44] Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., & Xia, W. (2021). Learning self-consistency for deepfake detection. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 15023–15033). IEEE.

[45] Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Chen, D., … Guo, B. (2020). *Identity-driven deepfake detection.* https://doi.org/10.48550/arXiv.2012.03930

[46] Jafar, M. T., Ababneh, M., Al-Zoube, M., & Elhassan, A. (2020). Forensics and analysis of deepfake videos. *2020 11th international conference on information and communication systems (ICICS)* (pp. 53–58). IEEE.

[47] Bondi, L., Cannas, E. D., Bestagini, P., & Tubaro, S. (2020). Training strategies and data augmentations in cnn-based deepfake video detection. *2020 IEEE international workshop on information forensics and security (WIFS)* (pp. 1–6). IEEE.

[48] Hongmeng, Z., Zhiqiang, Z., Lei, S., Xiuqing, M., & Yuehan, W. (2020). A detection method for deepfake hard compressed videos based on super-resolution reconstruction using CNN. *Proceedings of the 2020 4th high performance computing and cluster technologies conference & 2020 3rd international conference on big data and artificial intelligence* (pp. 98–103). Association for Computing Machinery.

[49] Han, J., & Gevers, T. (2020). *Mmd based discriminative learning for face forgery detection* [presentation]. Proceedings of the Asian conference on computer vision (pp. 1-17). http://openaccess.thecvf.com/content/ACCV2020/html/Han_MMD_based_Discriminative_Learning_for_Face_Forgery_Detection_ACCV_2020_paper.html

[50] Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. (2020). On the detection of digital face manipulation [presentation]. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5781–5790). IEEE.

[51] Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). *Use of a capsule network to detect fake images and videos*. https://doi.org/10.48550/arXiv.1910.12467

[52] Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Capsule-forensics: using capsule networks to detect forged images and videos. *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2307–2311). IEEE.

[53] Rana, M. S., & Sung, A. H. (2020). Deepfakestack: a deep ensemble-based learning technique for deepfake detection. *2020 7th IEEE international conference on cyber security and cloud computing (cscloud)* (pp. 70–75). IEEE.

[54] Bonettini, N., Cannas, E. D., Mandelli, S., Bondi, L., Bestagini, P., & Tubaro, S. (2021). Video face manipulation detection through ensemble of CNNs. *2020 25th international conference on pattern recognition (ICPR)* (pp. 5012–5019). IEEE.

[55] Chintha, A., Thai, B., Sohrawardi, S. J., Bhatt, K., Hickerson, A., Wright, M., & Ptucha, R. (2020). Recurrent convolutional structures for audio spoof and video deepfake detection. *IEEE journal of selected topics in signal processing*, *14*(5), 1024–1037.

[56] Masi, I., Killekar, A., Mascarenhas, R. M., Gurudatt, S. P., & AbdAlmageed, W. (2020). Two-branch recurrent network for isolating deepfakes in videos. *Computer vision–eccv 2020: 16th european conference, glasgow, uk, 2020, proceedings, part vii 16* (pp. 667–684). Springer.

[57] Tariq, S., Lee, S., & Woo, S. S. (2020). *A convolutional lstm based residual network for deepfake video detection*. https://doi.org/10.48550/arXiv.2009.07480

[58] Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., & Natarajan, P. (2019). Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, *3*(1), 80–87.

[59] Sohrawardi, S. J., Chintha, A., Thai, B., Seng, S., Hickerson, A., Ptucha, R., & Wright, M. (2019). Poster: towards robust open-world detection of deepfakes. *Proceedings of the 2019 ACM sigsac conference on computer and communications security* (pp. 2613–2615). Association for Computing Machinery.

[60] Güera, D., & Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)* (pp. 1–6). IEEE.

[61] Amerini, I., Galteri, L., Caldelli, R., & Del Bimbo, A. (2019). Deepfake video detection through optical flow based CNN. *Proceedings of the IEEE/CVF international conference on computer vision workshops* (p. 10). IEEE.

[62] Trinh, L., Tsang, M., Rambhatla, S., & Liu, Y. (2021). Interpretable and trustworthy deepfake detection via dynamic prototypes. *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 1973–1983). IEEE.

[63] Du, M., Pentyala, S., Li, Y., & Hu, X. (2020). Towards generalizable deepfake detection with locality-aware autoencoder. *Proceedings of the 29th ACM international conference on information & knowledge management* (pp. 325–334). Association for Computing Machinery.

[64] Nguyen, H. H., Fang, F., Yamagishi, J., & Echizen, I. (2019). Multi-task learning for detecting and segmenting manipulated facial images and videos. *2019 IEEE 10th international conference on biometrics theory, applications and systems (BTAS)* (pp. 1–8). IEEE.

[65] Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., & Verdoliva, L. (2018). *Forensictransfer: weakly-supervised domain adaptation for forgery detection*. https://doi.org/10.48550/arXiv.1812.02510

[66] Fernando, T., Fookes, C., Denman, S., & Sridharan, S. (2019). *Exploiting human social cognition for the detection of fake and fraudulent faces via memory networks*. https://doi.org/10.48550/arXiv.1911.07844

[67] Zhu, K., Wu, B., & Wang, B. (2020). Deepfake detection with clustering-based embedding regularization. *2020 IEEE fifth international conference on data science in cyberspace (DSC)* (pp. 257–264). IEEE.

[68] Lynch, S. S., Thigpen, C. A., Mihalik, J. P., Prentice, W. E., & Padua, D. (2010). The effects of an exercise intervention on forward head and rounded shoulder postures in elite swimmers. *British journal of sports medicine*, *44*(5), 376–381. https://doi.org/10.1136/bjsm.2009.066837

[69] Chiu, T. T. W., Ku, W. Y., Lee, M. H., Sum, W. K., Wan, M. P., Wong, C. Y., & Yuen, C. K. (2002). A study on the prevalence of and risk factors for neck pain among university academic staff in hong kong. *Journal of occupational rehabilitation*, *12*(2), 77–91. https://doi.org/10.1023/A:1015008513575

[70] Du, C. X. T., Trung, H. T., Tam, P. M., Hung, N. Q. V., & Jo, J. (2020). Efficient-frequency: a hybrid visual forensic framework for facial forgery detection. *2020 IEEE symposium series on computational intelligence (SSCI)* (pp. 707–712). IEEE.

[71] Cozzolino, D., Rössler, A., Thies, J., Nießner, M., & Verdoliva, L. (2021). Id-reveal: Identity-aware deepfake video detection. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 15108-15117). IEEE.

[72] Zhang, W., Zhao, C., & Li, Y. (2020). A novel counterfeit feature extraction technique for exposing face-swap images based on deep learning and error level analysis. *Entropy*, *22*(2), 249. https://doi.org/10.3390/e22020249

[73] Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). Emotions don't lie: an audio-visual deepfake detection method using affective cues. *Proceedings of the 28th ACM international conference on multimedia* (pp. 2823–2832). Association for Computing Machinery.

[74] Nirkin, Y., Wolf, L., Keller, Y., & Hassner, T. (2020). *Deepfake detection based on the discrepancy between the face and its context*. https://doi.org/10.48550/arXiv.2008.12262

[75] Yu, C. M., Chang, C. T., & Ti, Y. W. (2019). *Detecting deepfake-forged contents with separable convolutional neural network and image segmentation*. https://doi.org/10.48550/arXiv.1912.12184

[76] Parkin, D. M., Bray, F., Ferlay, J., & Pisani, P. (2001). Estimating the world cancer burden: globocan 2000. *International journal of cancer*, *94*(2), 153–156. DOI: 10.1002/ijc.1440

[77] Chai, L., Bau, D., Lim, S.-N., & Isola, P. (2020). What makes fake images detectable? understanding properties that generalize. *Computer vision–eccv 2020: 16th European conference, glasgow, uk, august 23–28, 2020, proceedings, part xxvi 16* (pp. 103–120). Springer.

[78] Ciftci, U. A., Demir, I., & Yin, L. (2020). How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals. *2020 ieee international joint conference on biometrics (IJCB)* (pp. 1–10). IEEE.

[79] Chang, X., Wu, J., Yang, T., & Feng, G. (2020). Deepfake face image detection based on improved VGG convolutional neural network. *2020 39th Chinese control conference (CCC)* (pp. 7252–7256). IEEE.

[80] Nguyen, H. M., & Derakhshani, R. (2020). Eyebrow recognition for identifying deepfake videos. *2020 international conference of the biometrics special interest group (BIOSIG)* (pp. 1–5). IEEE.

[81] Koopman, M., Rodriguez, A. M., & Geradts, Z. (2018). *Detection of deepfake video manipulation* [presentation]. The 20th irish machine vision and image processing conference (IMVIP) (pp. 133–136). https://www.researchgate.net/profile/Zeno-Geradts/publication/329814168_Detection_of_Deepfake_Video_Manipulation/links/5c1bdf7da6fdccfc705da03e/Detection-of-Deepfake-Video-Manipulation.pdf

[82] Baar, T., van Houten, W., & Geradts, Z. (2012). *Camera identification by grouping images from database, based on shared noise patterns*. https://doi.org/10.48550/arXiv.1207.2641

[83] Welch, B. L. (1947). The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, *34*(1–2), 28–35.

[84] Guarnera, L., Giudice, O., & Battiato, S. (2020). Deepfake detection by analyzing convolutional traces. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 666–667). IEEE.

[85] Agarwal, S., & Varshney, L. R. (2019). *Limits of deepfake detection: a robust estimation viewpoint*. https://doi.org/10.48550/arXiv.1905.03493

[86] Maurer, U. M. (2000). Authentication theory and hypothesis testing. *IEEE transactions on information theory*, *46*(4), 1350–1356. https://ieeexplore.ieee.org/abstract/document/850674/

[87] Hasan, H. R., & Salah, K. (2019). Combating deepfake videos using blockchain and smart contracts. *IEEE access*, *7*, 41596–41606. https://ieeexplore.ieee.org/abstract/document/8668407/

[88] Hasan, H. R., Salah, K., Yaqoob, I., Jayaraman, R., Pesic, S., & Omar, M. (2022). Trustworthy IoT data streaming using blockchain and IPFS. *IEEE access*, *10*, 17707-17721. https://doi.org/10.1109/ACCESS.2022.3149312

[89] Chan, C. C. K., Kumar, V., Delaney, S., & Gochoo, M. (2020). Combating deepfakes: multi-lstm and blockchain as proof of authenticity for digital media. *2020 IEEE/ITU international conference on artificial intelligence for good (AI4G)* (pp. 55–62). IEEE.

[90] Verdoliva, L. (2020). Media forensics and deepfakes: an overview. *IEEE journal of selected topics in signal processing*, *14*(5), 910–932. https://doi.org/10.1109/JSTSP.2020.3002101

[91] Dang, L. M., Hassan, S. I., Im, S., & Moon, H. (2019). Face image manipulation detection based on a convolutional neural network. *Expert systems with applications*, *129*, 156–168. https://doi.org/10.1016/j.eswa.2019.04.005

[92] Liu, Z., Qi, X., Jia, J., & Torr, P. H. S. (2019). *Real or fake: an empirical study and improved model for fake face detection* [presentation]. ICLR 2020 conference withdrawn submission (pp. 1-12). https://openreview.net/forum?id=HyxcZT4KwB

[93] Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019). *Protecting world leaders against deep fakes.* http://openaccess.thecvf.com/content_CVPRW_2019/papers/Media%20Forensics/Agarwal_Protecting_World_Leaders_Against_Deep_Fakes_CVPRW_2019_paper.pdf?source=post_page----------

[94] Battah, A. A., Madine, M. M., Alzaabi, H., Yaqoob, I., Salah, K., & Jayaraman, R. (2020). Blockchain-based multi-party authorization for accessing IPFS encrypted data. *IEEE access*, *8*, 196813-196825..

[95] Given, L. M. (2008). *The Sage encyclopedia of qualitative research methods*. Sage Publications.

[96] Maras, M. H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos. *The international journal of evidence & proof*, *23*(3), 255–262. https://doi.org/10.1177/1365712718807226

[97] Pitt, J. (2019). Deepfake videos and ddos attacks (deliberate denial of satire). *IEEE technology and society magazine*, *38*(4), 5–8. https://doi.org/10.1109/MTS.2019.2948416

[98] Couldry, N., & Hepp, A. (2013). Conceptualizing mediatization: Contexts, traditions, arguments. *Communication theory*, *23*(3), 191-202. https://doi.org/10.1111/comt.12019

[99] Hepp, A., Hjarvard, S., & Lundby, K. (2015). Mediatization: theorizing the interplay between media, culture and society. *Media, culture & society*, *37*(2), 314–324. https://doi.org/10.1177/0163443715573835

[100] Fotopoulou, A. (2016). Digital and networked by default? women's organisations and the social imaginary of networked feminism. *New media & society*, *18*(6), 989–1005. https://doi.org/10.1177/1461444814552264

[101] Potter, W. J. (2011). Conceptualizing mass media effect. *Journal of communication*, *61*(5), 896–915. https://doi.org/10.1111/j.1460-2466.2011.01586.x

[102] Rumpala, Y. (2012). Artificial intelligences and political organization: an exploration based on the science fiction work of iain m. banks. *Technology in society*, *34*(1), 23–32. https://doi.org/10.1016/j.techsoc.2011.12.005

[103] Dirican, C. (2015). The impacts of robotics, artificial intelligence on business and economics. *Procedia-social and behavioral sciences*, *195*, 564–573. https://doi.org/10.1016/j.sbspro.2015.06.134

[104] Olsher, D. J. (2015). New artificial intelligence tools for deep conflict resolution and humanitarian response. *Procedia engineering*, *107*, 282–292. https://doi.org/10.1016/j.proeng.2015.06.083

[105] Holder, C., Khurana, V., Harrison, F., & Jacobs, L. (2016). Robotics and law: key legal and regulatory implications of the robotics age (part i of ii). *Computer law & security review*, *32*(3), 383–402. https://doi.org/10.1016/j.clsr.2016.03.001

[106] Alharthi, R., Guthier, B., & El Saddik, A. (2018). Recognizing human needs during critical events using machine learning powered psychology-based framework. *IEEE access*, *6*, 58737–58753.

[107] Degerstedt, L., & Snickars, P. (2017). More media, more people—on social & multimodal media intelligence. *Human it*, *13*(3), 54–84.

[108] Schudson, M. (1989). The sociology of news production. *Media, culture & society*, *11*(3), 263–282. https://doi.org/10.1177/016344389011003002

[109] Kushin, M. J., Yamamoto, M., & Dalisay, F. (2019). Societal majority, facebook, and the spiral of silence in the 2016 us presidential election. *Social media+ society*, *5*(2), 2056305119855139. https://doi.org/10.1177/2056305119855139

[110] Janaszkiewicz, P., Krysińska, J., Prys, M., Kieruzel, M., Lipczyński, T., & Różewski, P. (2018). Text summarization for storytelling: formal document case. *Procedia computer science*, *126*, 1154–1161. https://doi.org/10.1016/j.procs.2018.08.053

[111] Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., … & Wu, Y. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, *31*. https://proceedings.neurips.cc/paper/2018/hash/6832a7b24bc06775d02b7406880b93fc-Abstract.html

[112] Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., … & Kreps, S. (2019). *Release strategies and the social impacts of language models*. https://doi.org/10.48550/arXiv.1908.09203

[113] Thies, J., Elgharib, M., Tewari, A., Theobalt, C., & Nießner, M. (2020). Neural voice puppetry: audio-driven facial reenactment. *Computer vision–eccv 2020: 16th European conference, glasgow, uk, august 23–28, 2020, proceedings, part xvi 16* (pp. 716–731). Springer.

[114] Kim, D., & Kim, S. (2018). Newspaper journalists' attitudes towards robot journalism. *Telematics and informatics*, *35*(2), 340–357. https://doi.org/10.1016/j.tele.2017.12.009

[115] Karnouskos, S. (2018). Self-driving car acceptance and the role of ethics. *IEEE transactions on engineering management*, *67*(2), 252–265. https://ieeexplore.ieee.org/abstract/document/8542947/

[116] Montal, T., & Reich, Z. (2017). I, robot. you, journalist. who is the author? authorship, bylines and full disclosure in automated journalism. *Digital journalism*, *5*(7), 829–849. https://doi.org/10.1080/21670811.2016.1209083

[117] Hopp, T., & Gangadharbatla, H. (2016). Examination of the factors that influence the technological adoption intentions of tomorrow's new media producers: a longitudinal exploration. *Computers in human behavior*, *55*, 1117–1124. https://doi.org/10.1080/21670811.2016.1209083

[118] Freidson, E. (2001). *Professionalism, the third logic: On the practice of knowledge*. University of Chicago Press.

[119] Hall, S. (1997). *Representation: cultural representations and signifying practices*. SAGE Publications.

[120] Bartneck, C. (2013). *Robots in the theatre and the media*. https://ir.canterbury.ac.nz/bitstream/10092/16697/2/bartneckDesForm2013.pdf

[121] Epstein, S. L. (2015). Wanted: collaborative intelligence. *Artificial intelligence*, *221*, 36–45. https://doi.org/10.1016/j.artint.2014.12.006

[122] Young, K. L., & Carpenter, C. (2018). Does science fiction affect political fact? yes and no: a survey experiment on "killer robots." *International studies quarterly*, *62*(3), 562–576. https://doi.org/10.1093/isq/sqy028

[123] MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction studies. social behaviour and communication in biological and artificial systems*, *7*(3), 297–337. https://doi.org/10.1075/is.7.3.03mac

[124] Mara, M., & Appel, M. (2015). Science fiction reduces the eeriness of android robots: a field experiment. *Computers in human behavior*, *48*, 156–162. https://doi.org/10.1016/j.chb.2015.01.007

[125] Clark, R. (2016). "Hope in a hashtag": the discursive activism of# whyistayed. *Feminist media studies*, *16*(5), 788–804. https://doi.org/10.1080/14680777.2016.1138235

[126] Lenhart, A., Ybarra, M., & Price-Feeney, M. (2016). *Nonconsensual image sharing: one in 25 Americans has been a victim of" revenge porn"*. https://apo.org.au/node/266206

[127] Karnouskos, S. (2020). Artificial intelligence in digital media: the era of deepfakes. *IEEE transactions on technology and society*, *1*(3), 138–147.

[128] Coeckelbergh, M. (2018). Technology and the good society: a polemical essay on social ontology, political principles, and responsibility for technology. *Technology in society*, *52*, 4–9. https://doi.org/10.1016/j.techsoc.2016.12.002

[129] Ancelovici, M., Dufour, P., & Nez, H. (2016). *Street politics in the age of austerity*. Amsterdam University Press.

[130] Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social media+ society*, *6*(1), 2056305120903408. https://doi.org/10.1177/2056305120903408

[131] Tufekci, Z. (2017). *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.

[132] Habermas, J. (2020). The public sphere: an encyclopedia article. In *Critical theory and society* (pp. 136–142). Routledge.

[133] Couldry, N., & Hepp, A. (2018). *The mediated construction of reality*. John Wiley & Sons.

[134] Gates, M. (2018). Is seeing still believing: factors that allow humans and machines to discriminate between real and generated images. *SMPTE motion imaging journal*, *127*(9), 70–78.

[135] Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of personality*, *88*(2), 185–200. https://doi.org/10.1111/jopy.12476

[136] Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: a survey on identification and mitigation techniques. *ACM transactions on intelligent systems and technology (TIST)*, *10*(3), 1–42. https://dl.acm.org/doi/abs/10.1145/3305260

[137] McGrew, S., Ortega, T., Breakstone, J., & Wineburg, S. (2017). The challenge that's bigger than fake news: civic reasoning in a social media environment. *American educator*, *41*(3), 4. https://eric.ed.gov/?id=EJ1156387